UNIVERZITA KARLOVA

Filozofická fakulta

Ústav anglického jazyka a didaktiky

DIPLOMOVÁ PRÁCE

BC. VERONIKA GIŽOVÁ

Types and use of shortening on Twitter

Typy a užívání zkratek na Twitteru

Praha 2017

Vedoucí práce: prof. PhDr. Aleš Klégr

Poděkování

Ráda bych poděkovala vedoucímu mé práce, prof. PhDr. Alešovi Klégrovi, za jeho vstřícnost, trpělivost a užitečné připomínky, jež mi věnoval v průběhu zpracování této diplomové práce. Rovněž děkuji mé rodině a přátelům, kteří mi po dobu psaní práce byli velkou oporou.

Prohlašuji, že jsem diplomovou *práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.*

V Praze dne 8. srpna 2017 ……..

ABSTRAKT

Východiskem práce je skutečnost, že komunikace na sociálních sítích, konkrétně na Twitteru, probíhá prostřednictvím krátkých textových zpráv, tweetů, které mají délku omezenou na maximálně 140 znaků. To vede k přirozené tendenci zkracovat jednotlivá slova, ale i víceslovné výrazy ve snaze ušetřit místo a zvětšit objem zasílané informace. Práce zkoumá hypotézu, že počet zkratek na Twitteru a jejich rozmanitost může sloužit jako stylistický indikátor tweetového žánru. Předpokládá se, (i) že počet zkratek a jejich typů v twitterovém vzorku bude vyšší než v jiných žánrech a (ii) bude obsahovat širší zastoupení metod krácení, z nichž některé budou příznakové pro tweetový žánr obecně v porovnání s kontrolním vzorkem. Zkoumaný vzorek 200 zkratek byl sebrán ze dvou twitterových trendů, *#Grenfell Tower* a *#Wimbledon*. V analýze vzorku je shromážděný soubor zkratek popsán kvantitativně, porovnán s kontrolním vzorkem a poté samostatně interpretován kvalitativně. Kompletní tabulka s vysvětlením zkratek je uvedena v příloze.

Klíčová slova: Twitter, zkratky, stylistika, internetová lingvistika, jazyk sociálních médií

ABSTRACT

The thesis works with the fact that communication on social network sites, particularly on Twitter, occurs in short text messages, tweets, which are restricted to the maximum of 140 characters. This leads to the tendency to shorten single and multiword expressions in order to save space and increase the content of sent information. The thesis examines the hypothesis that the number of shortenings on Twitter and their variation may function as a stylistic indicator of tweet genre. It is expected (i) that the number of shortenings in the Twitter sample will be higher compared to other genres and (ii) that the sample will contain more types of shortening, some of which will be characteristic for the tweet genre in general in comparison with the control sample. The research sample of 200 shortenings was collected from two Twitter trends, *#Grenfell Tower* and *#Wimbledon*. In the analysis part, the collected sample of shortenings is examined quantitatively, compared to the control sample and separately, interpreted qualitatively. The Appendix contains the complete table listing the meanings of all shortenings.

Key words: Twitter, shortening, stylistics, internet linguistics, language of social media

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

API – Application Programming Device

DM – Direct message

GT – Grenfell Tower

RT – Retweet

TA – Twitter Archiver

W – Wimbledon

LIST OF TABLES

1. INTRODUCTION

The present thesis examines the distribution of shortenings in online communication with focus on the microblogging site Twitter. The social network site was selected for the research because of its unique feature, limiting all submitted posts to 140 characters. This prompts the users to shorten some expressions in order to increase the capacity of their messages. It is presumed that the number and variety of shortenings found in the Twitter sample may function as a stylistic indicator distinguishing the tweet genre from other genres. The results are expected to show whether (i) the concentration of shortenings and the number of their types is higher in the tweet genre in contrast to other genres and (ii) whether there are any types of shortening found only within the Twitter sample in comparison to a control sample which would function as a stylistic indicator of the tweet genre in general.

The theoretical part of the thesis describes the types of shortening found in the primary sources. Since the available literature concentrates predominantly on word-formation processes, the presented types are blending, clipping and initialisms respectively. The focus then shifts to social network sites, offering a brief description of online communities and characteristics of the microblogging service Twitter. Next inspected is the emerging field of Internet linguistics which is concluded with findings of previously conducted researches on Twitter and the occurrence of shortenings in online communication.

The research part consists of two chapters. Methodology describes the parameters for collecting tweets which comprise the Twitter sample. For the research, 200 tokens of shortening were extracted by Twitter Archiver from the trending hashtags *#GrenfellTower* and *#Wimbledon*. The second part then presents a quantitative analysis of the sample, measuring the number and types of shortening with regard to the number of words and tweets in the Twitter sample and compares them with the control sample consisting of newspaper articles reporting on the topic of the hashtags. A qualitative analysis follows, presenting the shortenings in more detail concerning their types.

2. THEORETICAL BACKGROUND

2.1 WORD-FORMATION PROCESSES

The category of shortenings tends to be underrepresented in standard books on word-formation since shortening does not fall under the mainstream or regular word-formation focused on morphematic processes. Non-morphematic processes refer to means of creating new words which: "use at least one element which is not a morpheme" (Fandrych, 2008: 107). An example of a non-morphematic shortening is the production of an initialism. Only the initial letter of a word or words in a phrase is retained while the rest of the word is clipped. The initial cannot be characterized as standard morpheme, the smallest grammatical unit of meaning (Fandrych, 2008: 106). Furthermore, the process of element deletion varies from one shortening to another, thus the non-morphematic word-formation is classified as irregular and non-transparent. Some linguists such as Plag (2012: 13) avoid delving into the status of non-morphematic processes due to their supposed lack of productivity or difficult categorization, others like Marchand (1969: 452) treat certain shortenings such as acronyms and abbreviations as products of "word-manufacturing", a practice in which parts of words are combined to create "artificial new words".

Due to the inconsistency of terminology in the available literature, it is necessary to distinguish the types of shortening presented in this thesis. The general label shortening subsumes all types of non-morphemic word-formation processes as suggested by Cannon: "the common term shortening as the name of the division that produces blends, acronyms, abbreviations, and other reduced items" (Cannon, 1989: 106-7). The categories of shortening that will be further examined are as follows: blends, clippings and initialisms which comprise of acronyms and abbreviations. The cover term initialisms was adopted from Cannon and Bauer and Huddleston since acronyms and abbreviations are very similar.

2.1.2 BLENDING

Blends, also called portmanteau words (from the French 'portmanteau' meaning 'suitcase opening into two equal parts'), are characterized as "a sequence of two bases with reduction

of one or both at the boundary between them, as in *brunch* from *breakfast + lunch*" (Bauer and Huddleston, 2002: 1636). The reduction of the base results in a creation of a splinter which is combined with another splinter or a whole word. The term splinter was originally introduced by Berman (1961: 279) and later described by Adams (1973: 142) as a mostly irregular form that is neither a morpheme, nor a compound-element, although sometimes it may carry a meaning of a regular word which contains the splinter.

Despite the process being called unpredictable, there are several options on how to combine the splinters in order to create a blend. The distinction of four main classes was adopted from Bauer and Huddlestone (2002: 1636) as the most comprehensive:[1]

i. "The blend consists of the first part of the first base + the whole of the second base:

paratroops (parachute + troops) *telebanking (telephone + banking)*

ii. [The blend] consists of the whole first base + the final part of the second:

breathalyser (breath + analyser) *newscast (news + broadcast)*

iii. [The blend] consists of the first part of the first base and the final part of the second:

heliport (helicopter + airport) *stagflation (stagnation + inflation)*

iv. The central part is common to the two bases: there is overlap between them. In some cases there may be overlap in writing but not in speech (*smog*, /smDg/, from *smoke + fog*, /smouk + fog), or in speech but not in writing (*ballute*, /bslu:t/, from *balloon + parachute*, /baluin + paeraluit /:[2]

motel (motor + hotel) *sexploitation (sex + exploitation)*"

Plag (2012: 125) adds two more restrictions to the structure of blends that distinguish them from other shortening processes. According to him, the blends merge splinters on a syllabic level, whereas initialisms combine only initial letters and clipping does not undergo any amalgamations of parts. Blends also retain the length of the original words, particularly of the second constituent as is best illustrated by Bauer and Huddleston's type iv. mentioned above which is slightly extended due to the inclusion of two complete elements such as in *sexploitation* from *sex* and *exploitation*.

[1] The definitions and examples are separated in the original text.
[2] The brackets are missing in the original version as well.

Fandrych (2008: 113) presents a comprehensive table including even the rarer types and comments that most blends originate in oral medium with the exception of graphic blends which make sense only after inspecting the written form:

1.	initial and final splinter with overlap	affluenza - affluence + influenza, smog – smoke + fog
2.	two initial splinters with overlap	modem - modulator + demodulator
3.	two final splinters with overlap	Kongfrontation - King Kong + confrontation
4.	overlap of full words	thinspirations - thin + inspiration
5.	initial splinter + full word with overlap	emoticon - emotion + icon
6.	final splinter + full word with overlap	netiquette - internet + etiquette
7.	full word + final splinter with overlap	adultescent - adult + adolescent
8.	insertion of one word into the other	un-bloody-believable
9.	more than two constituents	Clinterngate - Clint + intern + gate
10.	graphic blends	shampagne - shame + champagne, royoil - royal + oil

Table 1: Types of blends[3]

According to Stockwell and Minkova (2001: 7): "blending is an area of word formation where cleverness can be rewarded by instant popularity" which is further discussed by Quirk et al. (1985: 1583) who discuss blends in commercial coinages: "where many types of neologisms are criticized adversely […], blends seem rather to be enjoyed". However, Quirk et al. also claim that this is the reason why blends are so short-lived since new blends tend to be restricted to particular products' slogans or they emerge in news headlines during a heavily publicised event only to be forgotten afterwards. Only a small number of splinters survive the marketing sphere and journalism to enter the general vocabulary such as *-gate* marking infamous affairs after the Watergate scandal producing neologisms such as *Clinterngate* or *Muldergate* or *-oholic* with its variation *-aholic* denoting an addiction, e.g.: *workaholic*, *chocoholic*, *shopaholic* (Bauer and Huddleston, 2002: 1137). In spite of these limitations, blending has been increasingly popular since the latter half of the 20th century and remains as one of the top word-forming processes alongside initialisms owing to its coinage of new technological terms and use in electronic communication (Fandrych, 2008: 111).

[3] The examples were altered to include explanations of the expressions.

2.1.3 CLIPPING

Clipping, also known as truncation, is a word-forming process in which a part of a polysyllabic word (or a multiword expression) is reduced often to a single syllable (cf. *zoo*) while the meaning is maintained and the shortened form remains in the same word class as the original (Bauer, 1983: 233). Quirk et al. (1985: 1580) further remark that the process mainly involves a change from the stylistic perspective, the shortened form shows the user's familiarity with the term, thus rendering it informal, casual, e.g.: *photo* from *photograph*, *mag* from *magazine*. In the case of referring to a sensitive material, the meaning is 'obfuscated' as is the case of a seemingly innocent girl name *Mia* used by people suffering from the mental disease *bulimia* on online forums (Fandrych, 2008: 114).

Regarding the truncated part of the word, Marchand (1969: 441, 446) notes that the process does not remove a morpheme since clipping exceeds morpheme boundaries, yet rather an "arbitrary part" that can be supplied by the speaker at any time. He concludes that clipping is heavily based on speech as there is only a small number of clippings related to spelling such as *zoo* from *zoological garden* and thus clipping does not relate to the prevalent grammatical word-formation techniques. Bauer and Huddleston (2002: 1634) use the term 'surplus' for the part that is removed and 'residue' for the part that remains. Fandrych (2008: 114) names the 'arbitrary part' a 'free splinter' and compares the process of detachment to splinters in blending. It is possible to speak of clipping in both processes for both techniques show lack of consideration for morpheme structure, syllable structure and stress placement. Whereas the splinter needs to be reattached to a new word element, the clipped free splinter enjoys an independent status.

Clipping may occur in four positions with the most frequent type being back-clipping or final-clipping. Fore and back-clipping is the least frequent to the point of rarely occurring. The following definitions were again borrowed from (Bauer and Huddleston, 2002: 1635):

i. "Back-clippings: surplus removed from the back, i.e. word-final, part of the original:

coke (cocaine) *doc (doctor)* *lab (laboratory)*

ii. Foreclippings: surplus removed from the front:

bus (omnibus) *cello (violoncello)* *phone (telephone)*

iii. Ambiclippings: surplus removed from both beginning and end:

flu (influenza) *fridge (refrigerator; BrE)* *tec (detective; BrE)"*

Some of the common truncated expressions are the result of clipping with the long, original form either lost or at least not immediately coming to mind such as *pantaloons* preceding *pants, wig* from *periwig* or because the resulting free splinter has an obscure, ambiguous or field-specific meaning such as *loot* for *lieutenant*, *brolly* for *umberella*, *con* meaning *confidence trick*, *convict* or *conductor* depending on the supplied context or slang (see also Marchand, 1969: 441, 447). Even individual words in multiword expressions may be truncated and combined such as *elin* from *electronic intelligence* or *kidvid* from *kid's video* to create a 'clipping compound' (Bauer and Huddleston, 2002: 1935).

Proper names, first names in majority, undergo clipping quite frequently – the speakers tend to use shortened familiar names more than the original: *Ben* from *Benjamin*, *Liz* from *Elizabeth*, *Tina* from *Christina*, *Sac* from *Sacramento* (Marchand, 1969: 441-45). Among the other proper names which can be clipped are surnames, e.g.: *Mac* from *Macauly*, *Montie* from *Montgomery* and city names, e.g.: *Cin* from *Cincinnati*, *Philly* for *Philadelphia*. There is no set rule on what part of the name becomes the splinter, yet there is a strong tendency for the primary stressed syllable to be clipped, e.g.: *Belle* from *Arabella*, *Abe* from *Abraham*, *Xan* from *Alexandra* (Plag, 2012: 119). Another method would be to retain the first syllable, especially when the syllable also carries the primary stress, e.g.: *Alf* from *Alfred*, *Barb* from *Barbara*, *Bart* from *Bartholomew*.

Bauer and Huddleston (2002: 1936) and Plag (2012: 117) differentiate a special type of embellished clippings which are created when a free splinter receives a suffix which adds either a diminutive or jocular tone:

i. -y, -ie and -ies suffixes denote an endearing or diminutive expression, e.g.: *Mandy* from *Amanda*, *barbie* from *barbecue*, *rellies* from *relatives*, *sunnies* from *sunglasses*
ii. -er, -ers or -o suffixes mark familiarity or jocular expressions, e.g.: *rugger* from *rugby football*, *preggers* from *pregnant*, *journo* from *journalist*

While the majority of clipped words belong to the class of nouns, even adjectives may be shortened, although the forms are comparatively rare, e.g.: *awk* from *awkward*, *comfy* from *comfortable*, *legit* from *legitimate* (Marchand, 1969: 447). Some verbs may also yield to the process, yet the occurrences are even rarer than clipped adjectives and may be rather the result of a clipped noun whose free splinter resembles a verb, e.g.: *canter* from *Canterbury* or *tot up* meaning *sum up* from *total* (ibid.).

From all the non-morphematic shortening processes examined in this thesis, clipping may be designated as the purest shortening process since with the exception of its subtype embellished clipping, splinters created by the removal of material from their base do not have to be attached to another word element as in blending or initialisms, but may function independently in a sentence (Fandrych, 2008: 114).

2.1.4 INITIALISMS

The category of initialisms is a problematic one in word-formation theory due to the terminological confusion it is subjected to. It includes two types of shortening in which the words of a multiword expression are reduced to their initial letters and in speech the resulting form is either read as a word (and called "acronym") or spelled, in which case it is called "abbreviation, "alphabaetism" or "initialism" by various authors, but I will use the term "abbreviation" in the following. Since the boundary between these two types is very narrow, the same form can sometimes be both spelled and read as a word, the two processes are best subsumed under one common label. I will use the label initialisms for both acronyms and abbreviations as suggested by Cannon (1989) and Bauer and Huddleston (2002) and will explain their identifying features further on.

The creation of an initialism involves a great loss of material as in clipping (it is in fact a case of multiple clipping) and blending. However, the reduction is even greater as only the initial letter of each constituent is kept before the initials are combined in a new word, e.g.: *asap* from *as soon as possible*, *CIA* from *Central Intelligence Agency* (Plag, 2012: 126). An exception may be made to preserve an extra letter or to switch the order of letters to aid pronunciation of the new coinage or to create a homograph of an existing expression; moreover, function words tend to be omitted in the final product, e.g.: *ESPRIT* from *European Strategic Programme for Research and Development in Information Technology* or *MISHAP* from *Missiles High-Speed Assembly Program* (Fandrych, 2008: 109).

The initials are divided into two categories primarily based on their phonological properties, but according to research conducted by Cannon (1989: 116), also on their structure:

> "An **abbreviation** is an item created from one or two first letters of all or most of the 1-5 constituents of an existing item. Medial free forms and bound forms may be constituents, and the resulting shortening is pronounced letter by letter; […] an **acronym** is created from the

> first letter (and infrequently the second or even third letters) of all or most of the 3-9 constituents of an existing compound."[4]

The structural differences indicate that abbreviations also consist of single word shortenings while acronyms need at least three constituents. Regarding the former, the single word abbreviations are rarely mentioned in the literature. Bauer and Huddleston (2002: 1632;34) give examples of compounds that are treated as abbreviations such as *postcard* shortened to *pc* and *tuberculosis* to *TB*. However, these examples are considered departures from strict initialisms and are not further examined.

Occasionally, initialisms may be a combination of both spelling and word-like reading and then it depends on the interpretation of the author which label they choose in instances when the first letter is spelled as in an abbreviation, yet the rest of the word behaves as an acronym, e.g.: *VTOL* pronounced /ˈviːtɒl/ from *vertical take-off and landing*. Also, the pronunciation may not be immediately apparent as is the case of *ARVN* from *Army of the Republic of Vietnam*. The word pattern of the four consequent consonants would suggest reading it as an abbreviation, yet its second variation spelling *Arvin* indicates its status as an acronym (Cannon, 1989: 115-6).

Whereas blending and clipping originate in oral tradition, the category of initialisms is heavily based in orthography (Bauer, 1983: 238). It may be nicely exemplified by *PERT* meaning *Program Evaluation and Review Technique*. If the acronym were created with phonology in mind, the initial letters would retain the phonetic value of the long form, therefore *e* from *evaluation* would be pronounced as /ɪ/ and *r* would have to lose its rhotic status in British English. Instead of /pɜːt/ the proper pronunciation would be /pɪrt/ (ibid.). This is also evidenced by initialisms that have a graphic shortened form, yet when spoken, the whole word is pronounced or when concerning Latin, its English phrase equivalent, e.g.: *lb* stands for *pound*, *i.e.* is read as *that is* or *for example* (Quirk et al., 1985: 1582).

In addition, the influence of writing on initialisms is evident in the orthographic variation of some units such as *asap* which may be spelled both, in lowercase and uppercase, with lowercase possibly containing full stops after each constituent (Plag, 2012: 127). The use of punctuation and/or uppercase contributes to the reading of initialisms as abbreviations, although as the words become part of general vocabulary, they have a tendency to lose the full stops and change to lowercase unless they are proper names with an established form such as

[4] The original text did not contain highlighting in bold type.

FBI from *Federal Bureau of Investigation*, or may be written both ways *Unesco* or *UNESCO* from *United Nations Educational, Scientific and Cultural Organization* (Bauer and Huddleston, 2002: 1634).

While uppercase acronyms, although pronounced as words, are easy to distinguish in writing, some of the lowercase acronyms without punctuation marks may become so established in general English as words that the original lengthy phrase may be forgotten or its presence would not immediately evoke the acronym; this often the case with technical terms, e.g.: *radar* from *radio detection and ranging*, *laser* from *light amplification by the stimulated emission of radiation* (Bauer and Huddleston, 2002: 1634). Some instances even result in accidental reduplication of one of the shortened elements when a person uses the initialism as an adjective as in the phrase *PIN number* meaning *personal identification number number* (Fandrych, 2008: 110).

The source words of initialisms tend to be nouns of which some may be modified by adjectives which usually become represented within the shortening in contrast with grammatical words that only get included when they aid the overall form in pronunciation or help to give the form of an existing expression, e.g.: *FIST* from *Federation of Inter-state Truckers* versus *GRAS* from *Generally Recognized As Safe* (Bauer, 1983: 237). Regarding the behaviour of initialisms (both abbreviations and acronyms) in a sentence, the majority tends to function like common nouns, taking modifiers, plurals and possessives (Cannon, 1989: 109). Strangely, some collective nouns already denoting plural may have double variation in spelling, e.g.: *HQ* or *HQs* for *headquarters*. Once established in general vocabulary, the initialisms take part in further word-formation processes as illustrated by Fandrych (2008: 110):

Blending	InteracTV - Interactive + TV
(Multiple) Compounding	CD-Rom - Compact Disc Read-Only Memory
Conversion	to R.S.V.P. - to please respond from French *Répondez s'il vous plaît*
Prefixation	Un-PC - not politically correct
Suffixation	OK-ness - oll korrekt ness, something being fine

Table 2: Word-formation based on initialisms

Cannon (1989: 102) notes that there was a surge in the production of initialisms during World War Two when new military terms were needed and since then the process has remained largely productive in creating technological, scientific terms or names for institutions, places, programs, although many of them may only be relevant locally. Since then the initialisms may be found virtually anywhere, from corporations to news discourse, or they may be adopted privately as in-group slang expressions allowing the participants a certain amount of secrecy, fellowship through ironic intentions or jocular reasons, especially with the emergence of electronic communication and social media (Fandrych, 2008: 109-10, 115).

2.2 SOCIAL MEDIA

2.2.1 SOCIAL COMMUNITIES

Before the emergence of social network services, the content of the World Wide Web consisted mostly of information issued by commercial media or published by distinct individuals; the Internet was meant to be consumed as there was little opportunity for people to become content creators (Obar and Wildman, 2015: 746). With a shift to Web 2.0 applications and decrease in cost for online data storage, the Internet became interactive, changing the status of people from consumers to participants or "prosumers" a blend coined by Alvin Toffler to encompass both the consumer and producer (Ritzer and Jurgenson, 2010: 17). The applications enable internet users to create original content, interact with one another, collaborate on projects, modify existing material intended to be consumed and share virtually any information or data across the web due to social network services (Kaplan and Haenlein, 2010: 61).

Social network service, also known as SNS, social media or social networking[5] service is an internet platform used by individuals to engage in social interactions with other people on which they may freely express themselves and maintain interpersonal connections (boyd and

[5] boyd and Ellison argue against the use of 'networking' as the term implies that active search and engagement in social interaction with strangers is the predominant function of social media, resulting in a misleading emphasis (2007: 211).

Ellison, 2007: 211). Rather than forming new connections with strangers, the individuals tend to use SNS as a medium of preservation of the relations established in the offline world (Ellison, Steinfield and Lampe, 2007: 1155). The social networks vary in their content-orientation but there is a number of common features shared across the services: methods of registering a user profile, managing a network of connections and engaging in social interactions (Obar and Wildman, 2015: 746).

In order to access the service, the users need to create a public or a private profile under the terms and conditions of the respective owner companies by supplying identifying information about their online person – the requirements vary, yet usually include picking a username and a password, uploading a picture as their avatar and listing certain contact details. The supplied personal information is necessary, without it, the users would not be able to find and connect with other users on that particular site (Obar and Wildman, 2015: 747). Furthermore, some SNS sites, such as LinkedIn or Pinterest, do not allow anonymous viewers to access the contents without registration although the most popular sites give their users the option to select the level of privacy or the lack thereof in the settings, e.g.: Facebook, Twitter, or Instagram.

When the registration is completed, the new members are expected to create a network of connections with which they intend to interact. LinkedIn remains calling them ‘connections’ while other services such as Facebook and Snapchats opt for the familiar term ‘friend’ or the neutral ‘follower’ on Twitter or ‘subscriber’ on YouTube (Obar and Wildman, 2015: 747). Some social media applications have special coding; instead of a manual selection, the feed of posted information is viewed and shared by people within a particular geolocation, e.g.: Yik Yak application.

The level of social interaction depends on the service. Some applications offer a vast array of options such as Facebook that allows adding text, picture, video or audio posts, enables live streaming, lets the users manage thematic groups, organize events and most notably, it introduced the instant messaging application Messenger that placed it on the top of the list of most downloaded mobile applications in 2016 with 59.7 million downloads (McAlone, 2016). In contrast, the indie application Yik Yak targets their audience with a simple premise of adding anonymous text posts that are subsequently upvoted or downvoted by people in a restricted area around the original poster.

The social media have nowadays become an integral and inseparable part of everyday life since the new technology became the fastest means of reaching wide masses of people and the CEOs of large companies are not the only ones in charge of what is to be consumed anymore. Virtually anyone can share anything as long as they abide by the terms and conditions of the platform they choose. In the Information Age, the social media occupy an important part in society as the politicians and businesses use SNS to draw followers and customers, to speak for their actions and address a broad scope of issues to which people may readily react while also freely communicating with each other and discussing the current events – the information may be easily accessed from news outlets reporting on the events simultaneously as they are happening (Obar and Wildman, 2015: 747). In order to let this all unfold, people require language and with the new media, the users' languages are changing to accommodate the new needs of the online world which comes with its unique requirements such as the microblogging service Twitter with its 140-character limit per post. For linguistic research, the social media constitute a rich source of relatively free data that once harvested may prove invaluable for the future of linguistic research (Weller, 2014: 238, Zappavigna, 2011: 789).

2.2.2 TWITTER

The social service Twitter is a microblogging platform enabling the registered users to post short text messages called tweets roughly corresponding to thoughts or ideas (Russel, 2011: 7). Launched in 2006, the application first started as a side project under the Odeo company focused on podcasts but quickly rose to popularity after winning an award at *South by Southwest Interactive* 2007 conference. Twitter has become one of the top social media services nowadays since any major event gets immediately reported, shared and commented on by the community as attested on the day of 2016 U.S. presidential election when over 40 million tweets were sent on the topic, dominating the other social media (Isaac and Ember, 2016). Currently, Twitter has over 313 million monthly active users with over a billion of visits to sites with embedded tweets. Due to available smartphone technology, 82% of active users prefer to use the mobile applications over the Internet browser interface (About Twitter, 2017).

To understand the inner workings of Twitter, one needs to become acquainted with the basic terminology of Twitter communication familiar to all registered users. Table 3 provides a brief overview:

Direct Message (DM)	Private message sent from one user to another, not displayed with the other posts in the feed but in a separate tab
Followers	People who subscribe to other accounts
Following	People to whom one's account is subscribed
Hashtag (#)	The tag used to label a post which may be filtered through in the search engine
Like	All post may be 'liked' to show how many people agree with the content of the post
Mention (@)	In order to tag another user within the post, to ensure they get notified about the tweet, one needs to add at sign followed by their twitter handle
Retweet (RT)	Any already posted tweet may get retweeted by another account with a source link to the original poster, usually with RT preceding the text
Trend	Twitter provides a ranking list of the most popular hashtags or phrases at any moment, may be filtered through by location
Tweet	A Twitter post

Table 3: An Overview of Twitter terminology

The unique feature of Twitter is its policy to limit all posts to 140 characters, a concept based on mobile phones' instant messaging which has a similar restriction of 160 characters for the Roman alphabet. The reason behind the limitation is so that people using the service would be able to read the complete message at once even on their mobile devices[6] (Crystal, 2011: 36). The posted tweets chronologically accumulate on the author's personal page while simultaneously they are displayed in the feed of other people who chose to subscribe to the original poster's content and who may further react to the messages. Crystal (2012: 4) notes that the general thematic drive behind the messages switched after 2009 from "What are you doing?" to "What's happening?", making Twitter more news-oriented, focused on current events and fast reporting.

[6] The mobile phones circa 2006 had small screens since the device needed to reserve more space for the numeric keys which were abolished with the arrival of smartphones with touchscreens in late 2000s.

2.3 INTERNET LINGUISTICS

Many names have been given to the new field of linguistics focusing on language tendencies on the world wide web. Beginning in 1990s, the name then was *computer-mediated communication* but apart from linguistics, the term was too broad and included other forms of communication such as sending pictures, sharing music or video files. Furthermore, with more appliances being implemented with simpler versions of operational systems such as mobile phones or tablets, cutting loose from the traditional desktop computer gave rise to new potentially standard, yet broad terms *electronically mediated communication* or *digitally mediated communication* (Crystal, 2011: 1-2).

In order to relate to linguistics only, Crystal proposes to adopt *Internet linguistics* as the standard for there is a large number of various compounds of words containing *cyber/e/net/web* as their first part and *speak/lish/linguistics* as the second part, none of which have been established since "as a domain of academic enquiry, Internet linguistics is in its infancy" (Crystal, 2011: 3). To aid the new field, Crystal presents one of the first studies delineating the treatment and use of language online in his *Language and the Internet* (2006) and in its updated continuation *Internet Linguistics: A Student Guide* (2011) which will be the primary sources in this chapter. However, as Crystal remarks, the progress of electronic communication proves to be quite challenging to keep up with as exemplified by the content of his books which by the time they get published tend to lack information on the emerging new technologies:

> "By way of anecdotal illustration, the first edition of my *Language and the Internet* appeared in 2001: it made no reference to blogging and instant messaging, which had achieved little public presence at that time. A new edition of the book was therefore quickly needed, and that appeared in 2006. It included sections on the language of blogs and of instant messages, but it made no reference to the social networking sites, which had achieved little prominence, and certainly no mention of Twitter, which arrived in the same year. Linguistic studies of the Internet always run risk of being out of date as soon as they are written." (Crystal, 2011: 10-11)

2.3.1 The Internet as a medium

One of the problematic areas of online communication is how to treat the linguistic material on the Internet, as a written or spoken medium? Crystal (2011: 17) points out the duality of the relationship between the two media when related to the expressions one uses for the description of electronic communication: "[…] we talk about having an email 'conversation', entering a 'chat' room, and 'tweeting'. On the other hand, we talk about 'writing' emails, 'reading' web 'pages', and sending 'texts'".

The great amount of data available online swings from one extreme to another. The traditional attributes of written texts such as having spatial restrictions, being static, permanent and the author being physically distant from the reader or not even knowing the reader, may be to an extent observed in many online periodicals, literary archives or on news sites that function similarly to their offline counterparts (Crystal, 2011: 20). In contrast, the spontaneity of instant messaging is reminiscent of some features of speech – time restriction, dynamicity, transience and both of the participant are either present or know about each other's existence and identity. With blogging or passive interpersonal interaction on social media being caught in the middle of the spectrum as the inclination to writing or speech is bound to the individual preferences of people. Where one may adhere to the rules of grammar and construct elaborate sentences, another may send an email full of typographical errors and fragmented syntax. Not to mention that the stylistic choices vary with the thematic orientation of a platform. According to Crystal, while internet communication displays both types of media, overall, the language tends to be perceived as writing with tendencies to copy some features of speech (2011: 21). The following subsections shall give insight into the most specific differences found between the traditional media and electronic communication.

2.3.1.1 Differences with speech

Crystal (2011: 21-28) notes three major differences between speech and online communication: the lack of simultaneous feedback, the use of emoticons and the ability to be engaged in several conversations at once.

Focusing on the lack of simultaneous feedback first, Crystal stresses the importance of the listener as an active participant in a conversation. While the speaker talks, the listeners are supposed to react to the utterance to indicate they are paying attention to the subject matter

and show their subsequent thoughts and feelings by supplying vocalizations and through the use of mimicry and gestures. The speaker thus receives an instant feedback and in the case of ambiguity or misunderstanding, the situation may be immediately clarified. Crystal argues that the successive feedback is not as effective despite the fast replies in instant messaging as the participants may not feel the temporal restrictions and delay or reduce their responses which may happen in real life as well but on a smaller scale when compared to slower messaging such as sending emails or leaving a comment on a forum. It may be argued that in instances when one desires to have a proper face to face conversation, several applications allow audio and video exchange happening in real time such as Skype or Facetime. The lack of feedback could be also solved by the second feature – use of emoticons.

The commentary on emoticons may be traced to 1990s with description ranging from simple "pictographs" (Thompson and Foulger, 1996: 226) to "smileys" (Sanderson, 1993: 1) found predominantly in email exchange. Rezabek and Cochenour (1998: 201) define emoticons as "visual cues formed from ordinary typographical symbols that when read sideways represent feelings or emotions". With their inclusion in text, people may even subtly express irony or sarcasm which may lead to decreasing the number of confusing moments in online communication. Crystal admits that emoticons may assist in instances when one is pressed for time or is limited by space to send a quick response, however, when it comes to ambiguity, the emoticons prove as productive in causing misunderstandings as much as they prevent them (Crystal, 2011: 23-24). Nowadays, the emoticons are overshadowed by their Japanese variation called emojis. Attributed to Shigetaka Kurita, the emojis take the concept of emoticons, yet instead of punctuation marks, the end product is a tiny pixelated image (Blagdon, 2013). Whether ambiguous or not, people continue to enjoy them to the extent, they demand social media providers to enable the coding of applications to include more and more as evidenced by Twitter's announcement in which they released and open-sourced list of 872 emojis after receiving numerous requests (Twitter Blog, 2014).

The last specific feature relating to speech in online communication is the ability to participate in several conversations simultaneously. Whereas in real world, a person may only hope to be a part of two conversations and managing to process all the information shared, the number of online conversations in which one may engage is limited only by their memory and attention span (Crystal, 2011: 24). Since the messages may be read with a delay, people may switch effortlessly between tabs on the computer screen or between multiple applications on their phone.

2.3.1.2 DIFFERENCES WITH WRITING

Although Crystal (2011: 28-32) lists three main differences between online communication and writing: hypertextuality, permanence and multiple authorship, the two latter concepts are closely interconnected and thus will be presented together.

Hypertextuality is one of the most prominent features of the Internet. It functions as a transitional element which enables people to move from one site to another by one click on a hypertext link. Without it, the users would be restricted to the sites of which they know the web address and as a result, the online interaction would be severely limited. Similar feature can be found in traditional writing as well. Footnotes, bibliography and in-text references direct readers to other texts or places where they may gather additional information on the topic or check the presented facts. However, not all texts contain them as they are an optional feature. While the level of hypertextuality varies across sites, one thing is certain, they remain essential to keep the Internet a functioning network (Crystal, 2011: 28).

Whereas data online may be edited almost any time, traditional writing is restricted in this aspect. Crystal (2011: 29) points out that "a piece of text is static and permanent on the page". It is almost impossible to alter the text once it is printed or written down. Although some stationery supplies can erase ink or cover it, the editing can be spotted and in some cases easily reversed by scratching it out. On the Internet, the rules are more dynamic. Whereas some sites allow their users to add and edit comments or even the posts such as Wikipedia or Urban Dictionary, which rely on community input, others present content that may be only altered by the owner of the web page or not at all, as is usually the case with news reporting sites such as BBC or The Guardian which allow their journalists to edit online articles but not the archived ones (Crystal, 2011: 30).

The issue of editing plays an important role when it comes to authorship. This may be best exemplified with the previously mentioned sites which are created by users in a collaboration. The users may interact among each other to decide what the content will be like but often that is not the case as the users add and edit posts independently. Since people have different stylistic preferences and their idea of what is relevant varies as well as their spelling abilities, the multiple authorship may result in a heterogeneous text which erases the "physical identity of a text" (Crystal, 2011: 31) as it is difficult to ascertain what is left of the original text.

2.3.2 STUDIES ON TWITTER LINGUISTICS

As a popular social network, Twitter appears in several studies either as the subject or as the provider of valuable research material. Williams et al. (2013: 392) conducted a survey of available research papers featuring Twitter for which they managed to accumulate 1 557 studies, noting that "we are reaching a point where individual researchers will not be able to be familiar with all the literature published". Since various disciplines find interest in Twitter, the scope of the research ranges from political science to computer studies and most importantly for this thesis, also encompasses linguistics (Weller, 2014: 238).

Since the microblogging service provides almost endless data supply of people's opinions as new posts appear every second, the research often focuses on the content of messages, especially on evaluative language as people react to events (Saif, 2016). Zappavigna (2011: 789) introduces the term 'searchable talk' and instructs on how to use hashtags to discover what people think about certain topics and how to classify evaluative language. In fact, there are numerous studies which offer advice on how to use hashtags to filter through the material (Scott, 2015) or which describe methods to gather and process the tweets (Russell, 2011). The methodology of this thesis was influenced by Crystal's Twitter study (2011) accompanying the description of Internet Linguistics.

Regarding shortening on Twitter, only one study was found. Moehkardi (2016) examines the patterns and meanings of word-formation processes in online discourse, citing Twitter as one of the sources. The research includes acronyms (initialisms), clippings and blends. The results show that acronyms have the potential to become real words once they adopt lowercase and affixes, the prevalent pattern in clipping is back-clipping and that in blending, the first element tends to be back-clipped and the second fore-clipped. Overall, it appears that shortening on Twitter is an unexplored territory, thus the subsequent research may prove valuable in determining whether shortenings may be perceived as a stylistic indicator of the Twitter genre.

3. METHODOLOGY

3.1 SUBJECT OF STUDY AND SOURCES

The research material was extracted from the social network site Twitter. The microblogging site was selected as the basis for the research on types and use of shortenings because of its unique feature restricting the length of tweets – the posts shared by Twitter users. Similar to texting, the posts are limited by 140 characters, which encourages the users to shorten words for economical reasons and to increase the content value of their messages. Despite obligatory registration on Twitter before one may start posting, the information shared by the users cannot be perceived as factual. Virtually anything can be entered as one's username, full name, bio or location, thus the research focuses only on the content of messages, specifically on the word-formation processes involved in shortening and on shortenings as a stylistic marker.

The gathered material was manually searched for instances of shortening until 200 tokens were collected. To determine their type and full-length version, the shortenings were examined in the context of their tweet and any related tweets through replies on Twitter. To ensure that they were correctly interpreted, several online dictionaries were consulted: *Oxford English Dictionary*, *Cambridge Dictionary, Wiktionary* and *Urban Dictionary*. Although *Wiktionary and Urban Dictionary* are crowdsourced, which means that anyone may submit their definition of any word, they are one of the most up-to-date and most comprehensive dictionaries of English colloquial expressions available.

3.2 TWEET COLLECTION METHOD

The tweets were collected by Twitter Archiver (TA), an add-on available in Google Webstore for free. To use it, one needs to be signed in their Google and Twitter accounts, open a Google spreadsheet and authorize a link between the two applications. The add-on lets the user specify certain parameters for extracting tweets similarly to the advanced search option

available on Twitter.[7] After creating a search rule, the program starts collecting tweets on a new spreadsheet list and updates it every hour. As regards past tweets, TA can only retrieve posts less than a week old. Apart from the textual content of tweets, TA automatically extracts additional metadata such as tweet's time stamp, the user's Twitter name, full name, bio, number of follows, followers, retweets and likes. All tweets come with their special identification number that hyperlinks to the original post on Twitter simplifying any subsequent checks.

For this research, I had to create two search rules. Since the free version of TA allows for one rule at a time, the data had to be collected on different days. Only two parameters were specified: 'these #hashtags' with *GrenfellTower* and *Wimbledon* and 'none of these words' with *RT*. The retweets needed to be excluded to avoid repetition of posts. The following parameters were considered, tested and rejected:

a) **Written in**: the language recognition tool is not yet very efficient, possibly due to the limited length of posts. When restricted by a hashtag, it omits a massive number of tweets in its lists, thus I opted for manual separation of tweets written in other languages than English.
b) **Near this place**: this option separates tweets based on their geolocation. Since not all users have enabled tracking on their phones or computers, the data retrieval for specified hashtags was hindered. The option could work for a location with a large population density such as London, however the results would still yield foreign languages and the sample would be compromised by more probable user repetition – some users tend to tweet multiple times in a row while others once in a while.

3.3 MOTIVATION UNDERLYING THE COLLECTION METHOD: THE AIMS AND PURPOSE OF THE ANALYSIS

The decision to collect the material from two thematically different trends was motivated by the goal to capture as many different types of shortenings found on Twitter as possible. The two trends (*#GrenfellTower* and *#Wimbledon*) were chosen to be thematically unrelated and

[7] The reasons for not using the advanced search tool are listed in Section 3.6.

widely different for this purpose (two different trends are potentially a better source than one trend). The hypothesis is that the number and variety of shortenings may function as a stylistic indicator distinguishing the tweet/Twitter genre from other genres (comprising the control sample). Accordingly, the data analysis will examine the quantitative and qualitative distribution of shortenings relative to the number of words and tweets in the whole sample. The results are expected to show whether (i) the concentration of shortenings and the variety of their types is higher in the tweet genre than in other genres and (ii) whether there are any shortening types exclusively found on Twitter (in comparison with the control genre) that would function as a stylistic marker of the tweet genre in general.

Since a full-scale comparison between the tweet genre and the control genre would exceed the permitted length of the thesis only a small control test sample was used. As the control genre, I chose newspaper articles (5) from the *BBC* (2), the *Guardian* (2) and the *Telegraph* (1) covering the same events from the same time, i.e. 14 June 2017 and 10 July 2017. The articles come from multiple sources as it proved difficult to find more than one article per news reporting site covering the sports event. The control genre sample was collected by gathering the articles until the word count matched the total from Twitter sample. The complete control sample contains 6 126 words out of which 3 161 words in two articles are on the topic of Grenfell Tower incident, while 2 965 words in three articles relate to the Wimbledon match between Nadal and Müller. 49 tokens of shortening were extracted from the five articles.

The tweets tagged with *#GrenfellTower* were originally posted on 14 June 2017 from 19:59 while tweets about *#Wimbledon* were posted on 10 July 2017 starting at 21:57. The reason for varying collection time between the two hashtags was the subsequent decision to incorporate another hashtag and waiting for another trending topic that would ensure the heterogeneity of users. The gathered material was then manually processed as indicated in Section 3.4 and then searched for the first 100 tokens of shortening per hashtag. The total of 6 540 words was gathered in 433 tweets for the extraction of 200 tokens. 3 637 words in 228 tweets belong to *#GrenfellTower* subcorpus while *#Wimbledon* subcorpus consists of 2 903 words in 205 tweets.

3.4 TWEET ELIMINATION CRITERIA

Not all tweets collected by TA were eligible for examination. The extraction parameters managed to reduce the number of random, irrelevant tweets, yet the sample needed to be filtered manually before collecting 200 shortenings. Some of the tweets were eliminated according to the following criteria:[8]

a) **The tweet was written in a language other than English.** Since the thesis focuses on English shortenings, the default data sample needed to be cohesive. While *#GrenfellTower* managed to secure tweets mostly in English for the event concerned a tragedy, a burning high-rise building happening in London at that time, and was of interest especially to British nationals, *#Wimbledon* was a tennis sports event followed by people around the world, thus the number of non-English tweets was greater.

(1) *Un voraz incendio arrasó con la #GrenfellTower en #Londres #14Jun*
(2) *Уимблдон. #Мюллер побеждает Надаля со счетом 15-13 в пятисетовом марафоне https://t.co/1Bc6KBFZVn #Wimbledon #ATP*

b) **The tweet was a spam.** Despite being tagged by the hashtag, the tweet itself did not contain any information concerning the subject. Rather, the hashtag was employed by the user in attempt to reach wider audience to advertise a service or a product.

(3) *Which country has the best flag in the world? #FlagDay #uk #USA #Jamaica #Canada #NBAFinals #GrenfellTower #ENGvPAK*
(4) *ASK me HOW to Earn car and $3600 weekly WhatsApp me at +233209619943 #Sarothemusical #Shefzy_TetelaVideo Vamos Rafa Gilles Muller #Wimbledon*

c) **The tweet was an enumeration of mentions or trending hashtags.** Not only was impossible to find out what language from the context, the tweets were characteristic for having no relevant content except for ranking the trends.

[8] The following examples were left unedited as they appeared on Twitter.

(5) *Top 5: 1: #Wimbledon +10 2: #MondayMotivation -1 3: Ed Orgeron +8 4: #SECMD17 -1 5: Coach O +6*

(6) *1. #Blackfish 2. #NEDAUT 3. #GrenfellTower 4. #novarock 5. London 2017/6/14 19:57 CEST #trndnl*

d) **The tweet was empty except for the hashtag**. This category also relates to tweets containing an extra variation of the hashtag and/or emojis but no accompanying words. They were deleted as there was no content from which I could sample the shortenings or determine the language.

(7) *#GrenfellTower #GrenfellFire*

(8) *#Wimbledon* 🎾

e) **The tweet was a repetition of an already posted tweet**. In the instances that one person published the tweet multiple times or another person retweeted the original post without tagging the message RT to avoid the filter, only one instance was left in the sample. This seemed to predominantly apply to retweeted posts from news and sports accounts.

(9) *Muslims who were awake to begin their Ramadan fast were 'a lifeline' in #GrenfellTower via @HuffPostUK*

(10) *Nadal loses 15-13 in 5th set, Venus wins, top-ranked Kerber loses at #Wimbledon ... http://www.news-journalonline.com/sports/20170710/nadal-loses-15-13-in-5th-set-venus-wins-top-ranked-kerber-loses-at-wimbledon ... @Wimbledon*

f) **The tweet was deleted by 17 July 2017**. Several checks were conducted during the analysis to discover whether all tweets in the sample were still available on Twitter with the final check on 17 July 2017. The elimination applied to all tweets, whether containing a shortening or not. Only one instance of deletion altering the results was found. The shortenings within the post were removed from results due to the incomplete nature of the tweet.

(11) *Can't believe you're gone Yas. You were always smiling & had endless words of wisdom. My good friend. My heart is b… https://t.co/tyihZwEwU9*

3.5 TOTAL WORD COUNT CRITERIA

To obtain the correct total of words from which shortenings could have been extracted, certain tweets or parts of tweets needed to be moved or deleted from the spreadsheets. The reasons for each elimination are explained below.

a) **All links were deleted**. The links either belonged to a picture, gif or a video shared as a reaction to the attached tweet or hyperlinked to another website. In the former case, the link was added to the tweet automatically by TA and would not be part of the message when viewed on Twitter. Since they have no content value and only function as hyperlinks, their inclusion would be misleading for they appeared in the majority of tweets.

(12) *Aftermath of a tragedy: Shocking scenes in London as emergency services search for fire victims #GrenfellTower https://t.co/ffqb6kT0VO*

(13) *Safety reviews are underway in the Black Country after the devastating #GrenfellTower fire in London today https://t.co/VVwqphMrCP*

b) **All emoticons and emojis were deleted**. While the use of emoticons was rare, tweets often included emojis. To determine their function and meaning would however be problematic for they generally lack any specific definitions and anyone may interpret them in number of ways. Extraction of shortenings would then be purely subjective. They should be rather treated as markers of social interaction online since they have a similar function of expressing feelings and showing reaction similarly to sharing pictures, gifs and videos (14) and (15). Furthermore, some emojis were not properly downloaded and/or supported by spreadsheets due to the large number of available emojis that increase in size with new updates and/or because of TA version 20 that was last updated on 13 June 2016. The emoji would then either appear as an empty rectangle as in example (16) or would not

be included at all as in (17) which on Twitter has an extra emoticon with the initials GB for Great Britain.

(14) *just caught up on the news for the first time today, I have no words. Rest in Peace you beautiful souls #Grenfelltower Such a tragedy :(*

(15) *@LionelMedia has me like* 🤔🤔🤔🤔🤔🤔🤔🤔🤔🤔 *all day long. #GrenfellTower*

(16) *#Wimbledon is a (quiet) class act for brands.* □🎾 *https://t.co/IJD7y4Tozu*

(17) *#GrenfellTower is gut wrenchingly sad BUT the way the #London people have pulled together and their generosity makes me PROUD* GB🚒🚑🚓

c) **Some mentions were deleted**. As the users may get engaged in twitter conversations by hitting the reply button, some of the downloaded tweets had incorporated mentions in the initial position. On Twitter, these mentions would not be part of the message, they would be placed above the tweet as metadata indicators of which users are engaged in the current conversation (18). The tweets thus had to be manually checked. In case, that users decided to tag another user within their message, the mention was left intact but moved to the final position of the tweet. The reason for the postponement was due to @ sign (19). Once a post with @ in the initial position was clicked, a spreadsheet function would activate and demand to replace the original letters viewed as error with correct cell identifiers.

(18) *@uk_chancellor @itvnews But at least he's been honest now. Honesty due from Barwell, Johnson & May over roles in #GrenfellTower?*

(19) *Why did you vote against making landlords ensure homes are fit for human habitation? #GrenfellTower #GrenfellFire #RESIST @Jesse_Norman*

In the later stage of research, it was decided to exclude shortenings occurring within usernames and focus only on the shortenings which were part of the user's message. While hashtags were employed to denote the user's feelings and opinions such as #WATTBA meaning 'what a time to be alive' (20) or #disgraceful (21) and thus qualified for shortening extraction, the mentions tended to be used as hyperlinks except for a few that occurred in apposition (22). The other obstacle was similar to the case of emojis. It proved difficult to find the meaning behind some of the shortenings within mentions as they could have been the

result of character restriction combined with the rule that two usernames cannot have identical form (23) and (24). In the end, the mentions were left as part of the final word count amounting to 6 540 words but were not drafted for shortenings.

(20) *Gilles Müller--Federer's throwback contemporary from the mid-Aughts--just beat Rafa Nadal, who is 4 years younger. #WATTBA #Wimbledon*
(21) *Finally, our so-called PM provides a statement. #disgraceful #Grenfelltower*
(22) *Sad to see so many people displaced by the #LondonFire at the #GrenfellTower !! I hope our PM @theresa_may will do she can for these people*
(23) *@ihtgw*
(24) *@RyanWJBCFC*

3.6 REJECTED METHODS

This section is intended as a warning to any future tweet sample collector as extracting tweets from Twitter proved to be rather challenging. While Twitter interface provides an advanced search tool which enables its users to filter through public posts based on a few options such as determining the hashtag, time stamp, geolocation and mentioned users, only a random sample is retrieved.[9] Furthermore, there is no option to download the tweets other than time-consuming manual selection. Apart from the search tool, the Twitter development team also offers access to their various application programming interfaces (API) to registered users. The disadvantage of this method of collecting tweets lies in that you have to be at least semi-proficient in the programming of several coding languages in order to create a custom application (Best Practices, 2017).

Before discovering TA, I planned to follow the 'mining recipes' from Russell's handbook (2011) on how to extract data from Twitter using the coding language Python. After several failed attempts, I concluded that the manual was intended for advanced Python programmers

[9] Since the time stamp only operates with days as the lowest unit of measurement, the search engine lists the currently popular posts on the top based on their likes and retweets. They remain unchanged in subsequent searches for some time until surpassed by other popular tweets while the rest of the tweets is randomized. Chronological ordering is not available yet.

who would be able to edit the search queries to better suit the subject of the study and eliminate software bugs. The same difficulties arose when working with the R coding language in the RStudio programme. The documentation available online was not helpful enough to form proper code strings. In addition, the tweets were downloaded with many orthographic errors such as empty rectangles signalling unsupported characters. In the end, I resorted to using TA which proved to be the best solution for the add-on also extracted metadata along with the content of tweets.

4. RESEARCH

4.1 DESCRIPTION AND ANALYSIS OF DATA

As mentioned in Section 3.2, the complete sample analysed in this chapter was collected by using two trends, the social event trend resulting in the #*GrenfellTower* subcorpus and the sports event trend yielding the #*Wimbledon* subcorpus. Each trend/subcorpus was searched for the first 100 tokens of shortening per hashtag. The complete sample combining both subcorpora includes the total of 6 540 words gathered in 433 tweets that were needed to extract 200 tokens of shortenings. Of these, 100 shortening tokens extracted from 3 637 words in 228 tweets belong to #*GrenfellTower* subcorpus while #*Wimbledon* subcorpus of another 100 shortening tokens consists of 2 903 words in 205 tweets.

The first part of the analysis focuses on the quantitative aspect of the research, examining the number of shortenings and their types in the Twitter corpus which is compared with the control sample consisting of the news articles described in Section 3.3. The second part of the analysis presents the data from a qualitative perspective, concentrating on specific representatives of shortenings. Since the examples in this section include only occurrences from the final version of Twitter corpus, the numbering of examples begins anew from (1).

In the process of classifying the shortenings, it was discovered that the sample contains 70 examples of shortening practices (35% of sample) that were not described in the primary literature which focused on word-formation processes. The unclassified items were thus grouped together based on similar formal features and further consulted with secondary literature. The resulting labels of “logograms”, “non-standard spellings” and “omitted letters” were sourced from a study on language of text messaging (Crystal, 2008). The category “complex shortenings” was devised for cases in which a combination of two or more processes of shortening was involved. A more detailed description of the types is available in Section 4.2.

4.2 Overall distribution in the Twitter corpus

4.2.1 Types and lemmas

Table 4 displays the overall distribution of shortening methods found within the complete Twitter sample. The extracted 200 tokens occur in 6 different categories and account for 3.1% of the total 6 540 words in the corpus.

Initialisms represent the most frequent shortening process with 86 tokens, comprising 43% of the whole corpus. The second place is occupied by logograms with 54 tokens (27%) and the third by clippings with 41 tokens (20.5%). Omitted letters and non-standard spellings add up to 9 and 7 tokens respectively, the former making 4.5% and the latter 3.5% from the whole sample. There were only 3 instances of complex shortening found in the corpus, accounting for 1.5% of the types of shortening. Despite mentioned as one of the basic shortening methods, blending lacks any representation in the collected sample and thus will not be further examined in Section 4.3 which focuses on the types of shortening in more detail.

Shortening processes	Total ∑	Total %	Lemma ∑	Lemma %
Blendings	0	0	0	0
Clippings	41	20.5	16	18.2
Complex shortenings	3	1.5	3	3.4
Initialisms	86	43	48	54.5
Logograms	54	27	8	9.1
Non-standard spellings	7	3.5	4	4.5
Omitted letters	9	4.5	9	10.2
Total of shortenings	200	100	88	100
Token-lemma ratio	0.44			
Non-shortenings	6340	96.9		
Shortenings	200	3.1		
Total of words	6540	100		

Table 4: The distribution of shortening types and lemmas in the Twitter sample

The data is further examined based on the number of distinct shortenings that are produced by the types. The last two columns in Table 4 display the total sum of lemmas found per

shortening process and their representation in percentages. It was decided against using the label type with regard to type-token ratio in order to avoid confusion since type is mainly used in the thesis in relation to practices of shortening. Instead, the tables contain the label lemma under which are subsumed the representatives of shortenings which vary in number, orthography or use of punctuation (cf. *QF* in ex. 1 and 2). The token-lemma ratio gives the number of unique shortenings (0.44) found in the sample.

(1) *#Wimbledon Men's* ***QF[10]*** *after today's play Murray v Querrey Cilic v Muller Raonic v Federer Berdych v Djokovic/Mannarino* (W 122)[11]

(2) *UPSET ALERT // Muller is through to the* ***QFs*** *after a 6-3 6-4 3-6 4-6 15-13 win over Nadal. He will face Cilic next. #Wimbledon* (A 57)

When the processes are ranked based on the unique lemmas they contain, the most productive shortening method remains the same – initialisms with 48 lemmas, accounting for 54.5% of all unique shortenings. However, the second place is no longer occupied by logograms but by clippings which include 16 lemmas (18.2%). Omitted letters follow with the exact same number of lemmas and tokens (9) but their representation in sample becomes higher (10.2%). Logograms are next with only 8 lemmas (9.1%) compared to their 86 tokens. Non-standard spellings constitute 4.5% of the unique lemmas with 7 items, while complex shortenings comprise 3.4% of the lemmas with 3 representatives, the same number as their tokens.

4.2.2 SHORTENINGS PER TWEET

The 200 tokens of shortening occur in 145 tweets, comprising 33.5% of all tweets as evidenced in Table 5. Approximately, every third tweet contains an example of shortening. Since the number of tokens is higher than the number of tweets they occur in, it was calculated that on average, one tweet[12] contains 1.38 shortenings.

[10] All following examples will be highlighted by bold.

[11] The information in parenthesis refers to the list of Twitter shortenings in Appendix (A) and to the assigned number of the shortening. Since only one example is offered per shortening, the second example was taken directly from the sample and the parenthesis refers to the excel document available with the online version of the thesis. The letter marks the list and trend (GT for *#GrenfellTower* or W for *#Wimbledon*) and the number marks the line.

[12] The tweet belongs to the category of 145 tweets containing shortenings.

Shortening representation in tweets	Total ∑	Total %
Tweets with shortening	145	33.5
Tweets without shortening	288	66.5
Total of tweets	433	100
Shortenings per tweet[13]	Total ∑	Total %
1	111	76.6
2	22	15.2
3	6	4.1
4	2	1.4
5	4	2.8
Total of tweets	145	100

Table 5: The overall distribution of shortenings in tweets

Most frequently, the tweets included only 1 shortening per tweet – 111 tweets altogether, accounting for 76.6% of the total 145 tweets. In 15.2% of the tweets (22), the messages contained 2 items. The frequency of occurrence dropped with each extra shortening per tweet as displayed in Table 5. The highest amount of shortenings found per tweet was 5, comprising 2.8% of the distribution. In total, there were 4 instances of such long tweets.

The slight deviation from direct proportion in decrease of frequency may have been caused by the character limit imposed on the users of Twitter and their struggle to incorporate more words within the tweet. To test this theory, it would be needed to compare the length of tweets in characters to the amount of shortenings included in them. It was decided against testing this approach because of the elimination of certain elements within the tweets in the initial stages of the analysis. The number of characters was altered, thus any calculations performed on them would not show proper figures. Moreover, the shortenings would need to be compared to their full-length versions to determine the number of characters lost, however, the unabridged wording of the shortenings is not certain in some cases.

[13] Since I needed only 100 tokens per hashtag, the last token in #*Wimbledon* was extracted from a tweet that contained two shortenings. Only the first one was used in the analysis but for the purpose of showing the proper number of shortenings per tweet, the tweet was classified as including two items.

4.2.3 COMPARISON WITH THE CONTROL SAMPLE

There were found 49 instances of shortening in the control sample which consists of 6 126 words. Compared with the result from Table 4, the shortenings comprise only 0.8% of the sample while the Twitter corpus contains 3.1% shortenings. The frequency of shortenings is thus almost 4 times lower in the control sample. Therefore, the high concentration of shortenings on Twitter indicates that shortenings can be taken as one of the most prominent stylistic indicators of Twitter discourse.

Shortening processes	Total ∑	Total %	Lemma ∑	Lemma %
Blendings	0	0	0	0
Clippings	4	8.2	2	9.5
Complex shortenings	1	2	1	4.8
Initialisms	38	77.6	15	71.4
Logograms	6	12.2	3	14.3
Non-standard spellings	0	0	0	0
Omitted letters	0	0	0	0
Total of shortenings	49	100	21	100
Token-Lemma ratio	0.43			
Non-shortenings	6077	99.2		
Shortenings	49	0.8		
Total of words	6126	100		

Table 6: The distribution of shortening types and lemmas in the control sample

The distribution of shortening processes in the control sample is lower in contrast with the Twitter sample. Table 6 shows that only 4 different types of shortening can be found in the control sample. While both samples lacked any example of blending, the control sample also lacked any instance of non-standard spellings or omitted letters, however, it contained 6 tokens of logograms (12.2%) which were not described in the primary literature. The inspection of the type in 4.3.4 should reveal the reasons why. Logograms appeared as the second most frequent type after initialisms with 38 tokens which comprise 77.6% of the control sample. Overall, initialisms seized the first place in both corpora, however, they operate as the primary shortening process in the control sample. In Twitter corpus, they comprise only 43% of the shortenings. The third place is occupied by clippings with 4 tokens. Their distribution among shortenings was lower than in the Twitter corpus since they have

only 8.2% share in comparison with 20.5%. The control sample also included one example of complex shortening (2%).

The examination of lemmas seems almost unnecessary in the control sample for the shortening processes are evenly represented. Initialisms remain as the most frequent type with 15 lemmas (71.4%), logograms follow with 3 lemmas (14.3%) and clippings are next with 2 lemmas (9.5%). Since the number of lemmas is lower, the single complex shortening now constitutes 4.8% of the sample. The lemma-token ratio seems almost identical in both samples. The control sample is only marginally less diverse with 0.43 than the Twitter corpus with 0.44. This leads to conclusion that while Twitter contains a higher number of shortenings, there is a high percentage of repetition among the shortenings.

Based on the observation that the logograms in Twitter sample consist of 8 lemmas only while their tokens are almost 7 times higher (54), it seems that the repetition concerns a select few. This is further evidenced in Table 7 which lists the top five lemmas from the Twitter sample. The most frequent lemma is the logogram *&* (A 1) with 33 tokens, comprising 16.5% of the sample containing 200 tokens. It is followed by the clipping *Rafa* with 21 tokens (10.5%). The high occurrence of *Rafa* (A 60) is easily explained. The subcorpus *#Wimbledon* focuses on the sports event in which *Rafa* (Rafael Nadal) played an important role. Thus, the shortening is occasion-specific. In the case of the ampersand, the repetition of the shortening seems as a result of being a popular space-saving device for it is represented in both subcorpora. In conclusion, the Twitter genre tends to contain a few selected shortenings with high frequency of distribution. These shortenings are either regularly used across trends such as the ampersand or are trend-specific such as *Rafa*.

TOP 5 SHORTENINGS			
Number	Shortening	∑	%
1	&	33	16.5
2	Rafa	21	10.5
3	u	9	4.5
4	vs	9	4.5
5	v	7	3.5

Table 7: Top 5 shortenings in the Twitter sample

4.3 SHORTENING TYPES

This section examines in detail the types of shortening that were found in the Twitter sample and compares them with the material described in the theoretical part and additionally with the classification of language of text messaging (Crystal, 2006). The aim of this analysis is to qualitatively assess the extracted shortenings from the Twitter sample and to determine whether the sample contains types of shortening that are stylistically characteristic for the tweet genre.

The complete distribution of the discovered lemmas is presented below in Table 8. Examples from both trends will be given to illustrate the methods of shortening. Since the shortenings require context in some cases to be properly interpreted, the examples will present the whole wording of the tweet message. The complete list of shortenings (lemmas) is available in Appendix which lists the meaning of items and one example of tweet per shortening.

Shortening	Examples	Lemma ∑
Clippings	ave, bro, champs, congrats, cray, Fab, gen, gent, inc, Ken, libdems, mins, Rafa, Regs, Tue, Wed	16
Complex shortenings	BldgRegs, Ken&C, w/in	3
Initialisms	am, apt, AO, ASAP, BBC, BST, CC, CET, CS, DM, ETA, etc, eu, GMB, ICYMI, KCTMO, LBC, lol, LMAO, mA, mm, Mr, nhs, ofc, OK, omg, pm, PM, PSA, QF, R, Rd, RD, rbkc, rip, SID, sm, SOAS, St, TL, TV, UK, UPS, US, v, vs, WATTBA, WTF	48
Logograms	&, @, £, 2, 4, K, r, u	8
Non-standard spelling	bcoz, cos, tho, wud	4
Omitted letters	as, av, bldg, hav, hrs, Rdbt, shld, smthing, tht	9
Total		88

Table 8: The distribution of shortenings within types

4.3.1 CLIPPING

Clipping occurs as the third most frequent method of shortening expressions in the sample with 41 tokens and 16 lemmas displayed in Table 9. The most frequent lemma is *Rafa* with 21 tokens, followed by *congrats* with 4 occurrences. Most of the lemmas (12) have only 1 representative in the sample.

CLIPPINGS							
Token ∑		41		Lemma ∑		16	
ave	1	cray	1	inc	1	Rafa	21
bro	2	Fab	1	Ken	1	Regs	2
champs	1	gen	1	libdems	1	Tue	1
congrats	4	gent	1	mins	1	Wed	1

Table 9: The distribution of clippings

Formally, the examples show the tendency to clip the final part – it is the only type of clipping represented within the corpus (ex. 3-4). It may be argued that *cos* is also an example of clipping (initial and final); however, based on Crystal's classification (2008: 48), this particular shortening falls under the category of non-standard spellings because of the alteration of vowels within the clipped version. If the item was spelled as *cause*, then it would be classified as initial clipping.

(3) *This must have been a day you and your fearless colleagues have truly been dreading* ***bro*** *#firefighters #GrenfellTower #EmergencyServices* (A 17)

(4) *#Nadal In such matches there are no losers A class match between 2 gentlemen The way sport should be played* ***Congrats*** *to both #Wimbledon* (A 22)

The majority of the clipped expressions were single words; the only exception was *libdems* (ex. 5). The multiword expression was shortened to the first syllable of each word similarly as is done in blending. Since both parts were initial and there was not observed any blending of splinters as described in Section 2.1.2, the compound is treated as a result of clipping.

(5) *Was today the best day to announce you were quitting #****libdems*** *@timfarron #GrenfellTower* (A 44)

There have been observed 6 instances of affixation among the clippings. 5 clippings retained their plural number and thus appear with *-s* suffix (ex. 6). Only was affixation process can be characterized as embellished clipping (ex. 7). *Cray* received the *-y* suffix which intensifies the evaluative adjective, especially since the shortening is followed by an exclamation mark.

(6) *Nothing wrong with Building **Regs** only the implementation of them #GrenfellTower* (A 65)
(7) *#Wimbledon is **cray**!* (A 24)

Some of the clippings only function as graphic shortenings (ex. 8-9). It is more plausible that *Tue* or *Wed* occur only in writing and when read aloud, the complete form *Tuesday* or *Wednesday* is pronounced instead. Since the research was conducted only with written examples, these are mostly speculations and it is possible that the shortened versions may be spoken in some slang, dialect or as a joke.

(8) #GrenfellTower - A40 closed both ways (no ETA for re-opening). Heavy traffic on all diversion routes, inc all inputs to Holland Park Rdbt. (A 37)
(9) Which is not to say that Rafa wouldn't either but odds on he would stand a better chance after the courts had baked on Tue/Wed #Wimbledon (A 76)

4.3.2 COMPLEX SHORTENINGS

COMPLEX SHORTENINGS			
Token ∑	3	Lemma ∑	3
BldgRegs		1	
Ken&C		1	
w/in		1	

Table 10: The distribution of complex shortenings

Complex shortenings are a minor word-formation process, represented in the Twitter corpus by 3 tokens only and the same amount of lemmas. This method was not explicitly described in primary literature and it may be argued that the items underwent the shortening separately and then were compounded, which would be the case of ex. (10-11). In this thesis, they are

treated like multiword clippings, with the difference that instead of one process, there are two or more shortening practices involved.

(10) *Fire regs & bldg control inspections are not fit 4 purpose. Update needed ASAP. #grenfelltower #**BldgRegs** #PartB* (A 16)

(11) *#GrenfellTower IS IT BECAUSE PRIME MINISTER U LOST **KEN&C** TO LABOUR SO WHY SHOULD U BOTHER U HORRIBLE HORRIBLE HUMAN BEING #troysout* (A 40)

(12) *Survivors said the one stairwell to escape **w/in** #GrenfellTower was allegedly blocked…* (A 84)

In example (10), *#BldgRegs* combine the method of omitted letters in *Bldg* with clipping in the second part *Regs*. Three shortening processes are involved in *KEN&C* (ex. 11). The first part *KEN* is clipped from *Kensington* and added to the initialism *C* standing for *Chelsea*. Both parts are then attached by the logogram &. Example (12) contains the shortening *w/in* which substitutes omitted letters with a dash. While the other instances were multiword expressions, *w/in* is the only example of a single word which went through two shortening processes (omitted letters and insertion of a logogram). The pronunciation of the complex shortenings remains uncertain for *BldgRegs* and *KEN&C* which could be either pronounced in its entirety or as *Building Regs* or *Ken and C*. In the case of *w/in*, the shortening functions as a graphic word and is pronounced as the regular long form *within*.

4.3.3 INITIALISMS

The initialisms occupy the first place as the major shortening process found in the Twitter corpus. They are represented by 86 tokens out of which 48 are lemmas. The most frequent initialism is *vs* meaning *versus* with 9 tokens and is closely followed by another variant of shortening *versus*, the single letter *v* with 7 tokens. The other more frequent lemmas are *CC* (5), *omg* (4) and *PM* (4), meaning *Centre Court*, *oh my god* and *Prime Minister*. Although there are 6 instances of the initialism *pm*, they are counted separately for two meanings are identified. First initialism represents the Latin phrase *post meridiem* translated in English as *past midday* while the second stands for *Prime Minister*. Most of the initialisms appear only once (30) as illustrated in Table 11.

INITIALISMS							
Token ∑			86	Lemma ∑			48
am	2	eu	1	OK	1	sm	1
apt	1	GMB	1	omg	4	SOAS	1
AO	1	ICYMI	1	pm	2	St	1
asap	2	KCTMO	3	PM	4	TL	1
BBC	2	LBC	1	PSA	1	TV	2
BST	1	LOL	2	QF	2	UK	2
CC	5	LMAO	1	R	1	UPS	1
CET	1	mA	1	Rd	1	US	2
CS	1	mm	2	RD	1	v	7
DM	1	Mr	2	rbkc	1	vs	9
ETA	1	nhs	1	rip	1	WATTBA	1
etc	1	ofc	1	SID	1	WTF	2

Table 11: The distribution of initialisms

The classification of initialisms proved to be the most challenging when it came to separating the shortenings into the subcategories of abbreviation and acronym. As it was already mentioned with clippings in 4.3.1, some of the shortenings seem to be bound to written medium only and additionally, should be labelled as graphic shortenings. In these cases, it is difficult to distinguish whether they belong among abbreviations or acronyms since the main distinction between the two subcategories lies in the pronunciation. The only structural difference that was noted in the theoretical part, concerned single words. Those are taken as abbreviations when they are clipped in a way that leaves only the initial letter or two letters. These abbreviations always operate as graphic abbreviations and when spoken, the entire word is pronounced.

There are only 2 lemmas that could be identified as pure acronyms based on the entries in online dictionaries (see 3.1): *rip* (ex. 13) and *SOAS* (ex. 14). The former stands for r*est in peace* from the Latin phrase *requiescat in pace* with the same meaning, the latter represents *The School of Oriental and African Studies*, a college of the University of London.

(13) *My thoughts on the #GrenfellTower catastrophe today... #**rip** #GlenfellTower #London* (A 66)

(14) *Why **SOAS** will always be home. #GrenfellTower* (A 71)

There are 3 initialisms that may be identified both as abbreviation and acronym for they can be spelled out or pronounced as a word: *asap*, *LOL* and *LMAO* (cf. ex. 15-16), respectively meaning *as soon as possible* /ˌeɪ.es.eɪˈpiː/ or /ˈeɪ sæp/, *lots of laugh* or *laughing out loud* /lɒl/ or /ɛləʊˈɛl/ and *laughing my ass off* /ɛlˌɛmˌeɪˈoʊ/ or /ləˈmaʊ/.

(15) *Forest Gate!!!! Donation transport **asap** #bedsforgrenfell #GrenfellTower* (A 10)

(16) *I'm running the #SanFranciscoMarathon in 2 weeks and I know how to inspire myself to the finish line now **LOL** #RafaNadal #Wimbledon #Nadal* (A 45)

(17) *Suddenly #Wimbledon looks so soo boring **LMAO*** (A 44)

There are two initialisms representing whole phrases *ICYMI* and *WATTBA* meaning *in case you missed it* and *what a time to be alive* respectively (ex. 18-19) which could be read as acronyms as well because of their distribution of vowels and consonants but there is no mention of their pronunciation in the dictionaries. Furthermore, while *ICYMI* has at least a brief entry explaining the meaning in almost all dictionaries, *WATTBA* is not mentioned anywhere else besides the *Urban Dictionary*. This suggests that they are not as widely used as the other initialisms and thus their pronunciation have not been standardized yet or they are pronounced as an entire phrase. In the case of the latter, the shortenings would function as a space-saving device in written medium. It can be also argued that when spoken as an acronym, the listeners would not recognize the shortening and mistake it for another word.

(18) ***ICYMI** | The big names all in action on a stunning day of tennis at #wimbledon #7tennis* (A 36)

(19) *Gilles Müller--Federer's throwback contemporary from the mid-Aughts--just beat Rafa Nadal, who is 4 years younger. #**WATTBA** #Wimbledon* (A 85)

The rest of the initialisms (41 lemmas) are either well-established abbreviations (ex. 20) or their mostly consonant pattern suggests they would be spelled out when spoken (ex. 21). In the case of single words among initialisms, they are subsumed under abbreviations by default (ex. 22). The shortening *bbc* stands for *British Broadcasting Corporation*, *CC* indicates *Centre Court* and *TL* means *timeline*.

(20) *If you're a journalist working in London and you're not out there asking tough questions about #GrenfellTower, why are you even there? #**bbc*** (A 13)

(21) *Mirka entering that **CC** stadium like #Federer #Wimbledon* (A 19)

(22) *Well that's one person on my **TL** who's happy, Clare Every cloud has a silver lining #Wimbledon* (A 75)

Although the single-word abbreviations were barely mentioned in the primary literature, they are quite numerous in the Twitter corpus for they comprise 10 lemmas out of 48 among initialisms or 88 from the whole sample and 25 tokens from 86 among initialisms or 200 from

the whole sample. Moreover, when I examined their structure, it showed that 5 of the 10 lemmas were created by medial clipping – only the initial and the final letter were preserved as in *Mr* standing for *mister* (ex. 23). There was one exception, *apt* representing *apartment,* which also retained the second letter (ex. 24). The other three abbreviations are *Rd* (A 63), *St* (A 72) and *vs* (A 83) meaning *road*, *saint* and *versus*.

(23) ***Mr** @joeottawaystyle at @wimbledon with mrporterlive in his Lock Monaco hat #Wimbledon* (A 49)

(24) *All that's left of #GrenfellTower London **apt** building 18 hrs after fire broke out. A painful reminder visible all around the neighbourhood.* (A 8)

Since the deletion of medial part appears in a pattern, the single-word shortenings could be also interpreted as a fourth type of clipping instead of initialism. However, it could be argued that the shortenings are subsumed under initialisms because the resulting item is not one or two syllables that were kept intact as in clipping. They are more reminiscent of initialisms created from compounds in which the initial letter of each component is preserved as in *mm* standing for *millimetre* (ex. 25). A more interesting research could be conducted in the future on the structural patterns of single-word initialisms to access whether they could be perceived as a separate category alongside abbreviations and acronyms. They already show distinct features when it comes to their structure and pronunciation. Structurally, there are three patterns of material deletion in the Twitter sample: medial (ex. 23-24), compound (ex. 25) and final (ex. 26) exemplified by *v* meaning *versus*. In terms of pronunciation, they are graphic shortenings and thus pronounced as their full-length version.

(25) *#paire looks about 2**mm** short of a radicalised beard #Wimbledon* (A 48)

(26) *#Wimbledon Men's QF after today's play Murray **v** Querrey Cilic **v** Muller Raonic **v** Federer Berdych **v** Djokovic/Mannarino* (A 82)

4.3.4 LOGOGRAMS

Logograms are the second most frequent shortening device, accounting for 54 tokens but only 8 lemmas. The ampersand & occurs as the most frequent logogram with 33 tokens, followed by the letter *u* with 9 tokens.

LOGOGRAMS			
Token ∑	54	Lemma ∑	8
&	33	4	2
@	3	K	1
£	2	r	1
2	3	u	9

Table 12: The distribution of logograms

Although they are included in the types of shortening, they do not behave as traditional shortenings. In word-formation, to create a shortening, a certain part of an expression must be deleted. In the case of logograms, no deletion happens. Instead, a symbol is used to represent a whole word or part of a word based on the pronunciation of the symbol (Crystal, 2008: 37). Thus, when the logograms are referred to as shortenings in this thesis, it means that they function as means of saving space (like traditional shortenings) but they cannot be assumed as a traditional shortening process in terms of word-formation such as clippings or initialisms.

The pronunciation plays an important role for it separates the logograms from pictograms which represent meanings of words with their visual shape. Crystal (2008: 38) mentions emoticons as an example of pictogram. In the preparatory stage of analysis, the emoticons and emojis were erased due to the subjective nature of their analysis, since there are no dictionaries to explain their meaning which even changes with context. Therefore, the Twitter corpus contains only examples of logograms, not pictograms.

To avoid confusion, it needs to be specified what logograms were considered as shortenings. Numerals were counted only when they operated as numeronyms meaning they represented a word homophonous with the numeral as in example (27). The numeral *2* appears in the sample in 3 instances with each token representing the preposition *to*. Although Crystal notes that logograms: “are part of the European ludic linguistic tradition” (2008: 41) that can be traced back several centuries, some linguists such as Borisova (2015: 7) claim they are more

frequent and characteristic for modern English. There were only 2 unique numeronyms in the sample, the second being *4* denoting the preposition *for* (ex. 28).

(27) *#borisjohnson should hang his head in #shame for his attitude **2** #london #fire #service & #Tory cuts #GrenfellTower* (A 4)

Single letters were counted as logograms when they represented a whole word because of their homophonous nature as can be observed in examples (28-29). The letter *r* stands for the verb *are* and the single letter *u* represents the pronoun *you*.

(28) *Substandard fire alarms & flammable materials in this day & age, terrible. #Kensington council **r** responsible **4** hiring #KCTMO #GrenfellTower* (A 58)

(29) *To those who have fallen in the #GrenfellTower.May you rest in peace to the hero's that keep going,we thank **u** from the bottom of our hearts* (A 78)

In the case of *K*, the classification proved difficult as there are two possible interpretations. The letter could be taken as an example of single-word initialism. As a graphic initialism, the word would be pronounced as *kilo*. In this context (ex. 30), it was decided that the letter functions as a logogram. Rather than pronounced as *kilo*, the letter would be spelt /keɪ/ when read aloud or the logogram would be read as *thousand*. Even though *kilo* stands for *thousand*, in the context of population count, *kilo* functions as a false synonym. Therefore, I would argue in favour of classifying *K* as a logogram in this context and as an initialism when it is pronounced as *kilo* (for example when it denotes weight).

(30) *Is Gilles Muller most famous person from Luxembourg after beating Rafa at #Wimbledon? Country has 570**K** population & not a single one I know.* (A 38)

The last type of logograms that was found in the collected sample concerns standard logograms. These standard symbols function similarly to the graphic shortenings mentioned with clippings and initialisms. In speech, the symbols are read as the word they represent. For instance, the ampersand (ex. 31), the most frequent logogram and shortening device that was found in the sample, evolved from a stylized form of the Latin conjunction *et* meaning *and* (*Wiktionary*). In tweets, the symbol often functioned as space-saving device.

(31) *So proud of the emergency services **&** community of London. Helpless yet still helping. We will keep praying. #pray #GrenfellTower* (A 1)

4.3.5 NON-STANDARD SPELLINGS

The non-standard spellings are a minor shortening process, occurring within the corpus 7 times in 4 unique lemmas. The most frequent representative of this type is *cos* with 3 tokens which also appears in the form *coz* with final consonant alternation.

NON-STANDARD SPELLINGS			
Token ∑	7	Lemma ∑	4
bcoz		1	
cos		3	
tho		2	
wud		1	

Table 13: The distribution of non-standard spellings

While Crystal (2008: 48) does not classify them as a shortening method since his category rather focuses on misspellings in general, in the Twitter sample, the non-standard spellings operate as graphic shortenings. The graphic form represents the pronunciation of the entire expression such as *bcoz* (ex. 32) which is read the same way as the full-length word *because* or in the case of *cos* (ex. 33), only the clipped part *cause*. *Cos* cannot be identified as a clipping due to the vowel alteration (*au* to *o*). The shortenings are rather attempts at phonetic spelling.

(32) *One of those games where u feel the outcome should have been decided by a coin toss* ***bcoz*** *neither deserved to lose #Wimbledon #nadal #muller* (A 14)

When it comes to *tho* (ex. 33), the shortening can be classified both as clipping or non-standard spelling. Since all instances of phonetic spellings were subsumed under non-standard spellings, *tho* was included in this category. The clippings *bro* and *Ken* are also spelt phonetically but while the non-standard spellings represent the whole word, clippings only represent the clipped part.

(33) *How amazing are our over stretched emergency services* ***tho*** *#GrenfellTower* (A 73)

4.3.6 OMITTED LETTERS

Omitted letters are quite low in numbers with only 9 tokens but high in lemmas since the figure is identical in both instances. As the label implies, omitted letters are created by character deletion (omission), mostly from the middle of the original word (Crystal, 2008: 45). The expression *bldg* is the result of removing all vowels and the consonant *n* from *building* (ex. 34)

(34) *Fire regs &* ***bldg*** *control inspections are not fit 4 purpose. Update needed ASAP. #grenfelltower #BldgRegs #PartB* (A 15)

OMITTED LETTERS			
Token ∑	9	Lemma ∑	9
as	1	Rdbt	1
av	1	shld	1
bldg	1	smthing	1
hav	1	tht	1
hrs	1		

Table 14: The distribution of omitted letters

Omitted letters are the third shortening method found in this sample that was not included in the primary literature. Their unpredictable creation is most likely the reason why. In contrast with other shortening processes such as clipping, it is difficult to assess what part of the word will be deleted. Mostly, the initial and final letters are preserved while the middle of the word is stripped of vowels and silent[14] or double consonants (Crystal, 2008: 46). This happened in example (34) but in the case of *as* and *av* (ex. 35-36), standing for *has* and *have* respectively, the initial consonant was removed and in the latter case, even the final vowel. The choice to employ these forms seems strange since the author could have used the contracted forms – *s* and *ve*. In the former case, it would even save an extra character.

(35) *Today* ***as*** *proved again Londoners will always be resilient and help others no matter who you are #GrenfellTower #Londoners* (A 9)

[14] By silent consonants are understood those consonants that are difficult to detect when the word is pronounced or are not pronounced at all.

(36) *From SID point of view, theres a need to rethink fire safety of high rise buildings. Suppression* ***av*** *proven fatally inadequate #GrenfellTower* (A 11)

Due to the lack of consistency, it is not possible in some cases to ascertain whether a shortening is a product of omitted letters or the author only misspelt it. For instance, the shortened form *tht* from *that* (ex. 37) occurs only once in the Twitter corpus but appears in several entries in the *Urban Dictionary* as an established shortening. To contrast it, the same tweet includes 2 additional items of omitted letters: *smthing* from *something* and *shld* from *should*. Whether the author was aware of the alternative spelling *tht* and modelled the other shortenings accordingly or whether all three items were created by accident cannot be found from the context. It can be only said with certainty that the items operate as shortenings, providing extra space for the rest of the words within the tweet.

(37) *Who even has a brain that works this way? Who'd vote against* ***smthing tht shld*** *be a basic right?Fucking @Conservatives #GrenfellTower #Tories* (A 74)

5. CONCLUSION

The aim of this thesis was to determine whether the number and the types of shortening found in the Twitter sample may function as a stylistic indicator of the tweet genre which would distinguish it from other genres. The hypothesis was tested from two perspectives. The first examined the extracted data quantitatively, meaning the frequency of distribution of the shortening processes and their variation were compared against the control sample. The second inspected the types of shortenings qualitatively to find out whether any unspecified types occur on Twitter, what they are and how they may be characteristic of the tweet genre.

The initial assumption that Twitter would contain a high number of shortenings was based on the fact that the social network site restricts their users to post text messages up to 140 characters only. The collected 200 tokens of shortenings were found in 145 tweets, in the total of 6 540 words. To find out if the number was of significance it needed to be compared with a control sample of similar size. The control sample consisted of 5 newspaper articles of the length of 6 126 words which reported on the same events as the Twitter users in the hashtags *#GrenfellTower* and *#Wimbledon* comprising the Twitter sample. The sample yielded only 49 shortenings which constituted 0.8% of the text. The distribution in the Twitter corpus was almost 4 times higher with 3.1% representation of shortenings. The results thus confirmed that Twitter contains a higher percentage of shortenings per word compared to the control sample. Concerning tweets, shortenings appeared in 33.5% of all tweets out of which 76.6% contained 1 instance of a shortening.

Apart from the distribution, the samples were also compared with regard to the types of shortening that were present in the corpora. The Twitter sample contained 6 various shortening processes while the control sample showed only 4 different types. Initialisms were the most frequent in both samples. They comprised the majority of the shortenings found in the control sample with 77.6%. It may be concluded that the control sample was mostly unified, showing one prevailing type of shortening while the other types were only marginally represented. The other three types were clipping, logograms and complex shortenings. All 4 types appeared in the Twitter sample as well; however, the distribution was more varied. Although initialisms occupied the first place when it came to the number of tokens and also lemmas, they constituted only 43%, less than a half of the sample. The other 57% was distributed among clipping, logograms, complex shortenings and two classes which appeared

exclusively in the Twitter sample, non-standard spellings and omitted letters. Logograms occurred more in the Twitter sample with 27% in contrast to 12.2% from the control sample. However, it was discovered that the high number of logograms was a result of a disproportionate amount of one shortening, the ampersand. The Twitter sample contained 33 instances of the conjunction & which comprised 16.5% of all shortenings. Similar situation repeated in clipping. The shortening Rafa constituted 21 tokens out of 41, and thus represented 10.5% of the Twitter corpus. The control sample displayed no such deviance, the number of lemmas was equal in the distribution of tokens. The distribution of lemmas among the tokens turned out to be almost identical with the Twitter sample showing the ratio of 0.44 and the control sample 0.43.

Overall, it may be concluded that while Twitter indeed contains a higher number of shortening types, the variation of the shortenings is similar to other genres. Since the results of the Twitter corpus were altered by the high occurrence of two shortenings, the ampersand and *Rafa*, a further study, examining the distribution of shortenings or perhaps only key words across thematic hashtags, may prove more insightful. It could determine what types of words tend to be tied to one trend and which appear consistently in all or in the majority of the trends. Such analysis would not be feasible with this Twitter corpus as it comprises of two hashtags only.

Before delving into the details of the qualitative analysis of the discovered shortening processes, it should be noted that one shortening method described in the primary literature lacked any representatives in the Twitter sample and also, in the control sample. There was no instance of blending; however, it cannot be said that the type is not productive. A further research would need to be carried out to find out whether the type is favoured in other genres or whether it appears on Twitter but under different circumstances, i.e. it may be thematically specific and thus occur only in certain trends.

The qualitative analysis further inspected the 6 types of shortening occurring in the Twitter sample. Only two of those types were described in the primary literature, clipping and initialisms. The shortenings were thus grouped together based on their similar features and consulted with additional sources. It was discovered that the shortening processes were reminiscent of the language of text messaging on the basis of which 3 categories were established: logograms, non-standard spellings and omitted letters. The category of complex

shortenings was devised to encompass those items that combined two or more shortening processes.

In the process of examinations, it was found out that while clipping could be easily identified in the analysis of shortenings with the prevalent type being back-clipping, initialisms were more problematic to determine. The subcategories of abbreviations and acronyms may be applied to established initialisms but when concerning the novel forms, the only certain identifying feature was that abbreviations subsume single words and phrases of two constituents while acronyms require at least three constituents. The aspect of pronunciation proved unfeasible in the analysis since plenty of the shortenings appear only in writing. Rather than sorting the initialisms into subcategories, it may be sufficient to label the uncertain shortenings as plain initialisms, especially concerning nonce words that come to be used only for a short amount of time before they are forgotten. This seemed to be the practise in *Cambridge Dictionary* which used the umbrella term abbreviation. In the case that one decides to distinguish the subcategories, I propose to view the single word abbreviations as a separate class. The analysis showed that the single words have a unique structure, usually retaining the initial letter or two initial letters and the final letter such as *Mr* standing for *Mister* or *apt* for *apartment*. In terms of pronunciation, these initialisms function as graphic shortenings and in speech are pronounced as whole.

Concerning the shortenings which were unspecified in the primary literature, the occurrence of non-standard spellings and omitted letters are characteristic of the tweet genre and thus can be classified as stylistic indicators of the tweet genre. However, they are not exactly an exclusive stylistic marker as the shortenings originate in the language of text messaging. Arguably, since the microblogging social network was modelled after texting, these shortenings can be perceived as stylistically inherent to both genres for they are closely interconnected.

Regarding the logograms, the reason for their lack of mention in the primary sources is most likely due to the fact that they are not a shortening word-formation process but rather a space-saving device. They were included in the analysis for they function in the same manner as the other shortenings, the only feature setting them apart is that while shortenings are created from their longer version, the logograms are symbols that represent the entire word. Since they occur in the Twitter sample as well as in the control sample, they are not considered characteristic of the tweet genre.

To address the research hypothesis, the thesis confirmed that Twitter contains a higher concentration of shortenings as well as a higher number of shortening types compared to other genres. Non-standard spellings and omitted letters were identified as the two most characteristic types which can function as stylistic indicators of the tweet genre but also of the language of text messaging after which the microblogging social network was modelled.

SOURCES AND REFERENCES

PUBLICATIONS

Adams, V. (1973). *Introduction to Modern English Word-Formation*. London: Longman.

Bauer, L. (1983). *English Word-formation*, Cambridge: Cambridge University Press.

Bauer, L. and Huddleston, R. (2002). "Lexical word-formation." In: Huddleston, R., Pullum, G.K., *The Cambridge Grammar of the English Language*, Cambridge: Cambridge University Press.

Berman, J.M. (1961). "Contribution on Blending." *Zeitschrift für Anglistik und Amerikanistik*, 9: 278-281.

Bloomfield, L. (1933). *Language*. New York: Holt.

Borisova, T. (2015). "Numeronym as a modern type of abbreviation in modern English." *Science and Education a New Dimension. Philology*, 70: 7-8.

boyd, d. m. and Ellison, N. B. (2007). "Social Network Sites: Definition, History, and Scholarship." *Journal of Computer-Mediated Communication*, 13.1: 210-230.

Cannon, G. (1989). "English Abbreviations and Acronyms in Recent New-Words Dictionaries." *American Speech* 64: 99-127.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.

Crystal, D. (2011). *Internet Linguistics: A Student Guide*. New York: Routledge.

Crystal, D. (2006). *Language and the Internet*. Cambridge: Cambridge University Press.

Crystal, D. (2012). "On myths and mindsets." *Lingua Montenegrina*, 2.10: 3-6.

Crystal, D. (2008). *Txtng: The Gr8 Db8*. Oxford: Oxford University Press.

Ellison, N., Steinfield, C., and Lampe, C. (2007). "The benefits of Facebook 'friends': Exploring the relationship between college students' use of online social networks and social capital." *Journal of Computer-Mediated Communication*, 12.3: 1143-1168.

Fandrych, I. (2008). "Submorphemic elements in the formation of acronyms, blends and clippings." *Lexis: Journal of English Lexicology*, 2: 105-123.

Haspelmath, M. (2002). *Understanding Morphology*. Oxford: Oxford University Press.

Kaplan, A. M. and Haenlein, M. (2010). "Users of the world, unite! The challenges and opportunities of social media." *Business Horizons*, 53.1: 59-68.

Marchand, H. (1969). *The Categories and Types of Present-Day English Word-Formation*, München: Beck.

Moehkardi, R. R. D. (2016). "Patterns and Meanings of English Words through Word Formation Processes of Acronyms, Clipping, Compound and Blending Found in Internet-Based Media." *Humaniora*, 28.3: 324-338

Obar, J. A. and Wildman, S. (2015). "Social media definition and the governance challenge: An introduction to the special issue." *Telecommunications Policy*, 39. 9: 745-750.

Plag, I. (2012). *Word-formation in English*, 9th ed. Cambridge: Cambridge University Press.

Quirk et al. (1985). *A Comprehensive Grammar of the English Language*, Longman.

Rezabek, L. L., and Cochenour, J. J. (1998). "Visual cues in computer-mediated communication: Supplementing text with emoticons." *Journal of Visual Literacy*, 18: 201-215.

Ritzer, G. and Jurgenson, N. (2010). "Production, consumption, presumption the nature of capitalism in the age of the digital 'prosumer'." *Journal of Consumer Culture*, 10.1: 13-36.

Russell, M. A. (2011). *21 Recipes for Mining Twitter*. Sebastopol, Cambridge: O'Reilly Media.

Saif, H. (2016). "Contextual semantics for sentiment analysis of Twitter." *Information Processing & Management*, 52.1: 5-19.

Sanderson, D. (1993). *Smileys*. Sebastopol, Cambridge: O'Reilly.

Scott, K. (2015). "The pragmatics of hashtags: Inference and conversational style on Twitter." *Journal of Pragmatics*, 81: 8-20.

Stockwell, R. and Minkova, D. (2001). *English Words: History and Structure*. Cambridge: Cambridge University Press.

Thompson, P. A., and Foulger, D. A. (1996). "Effects of pictographs and quoting on flaming in electronic mail." *Computers in Human Behavior*, 12: 225-243.

Walther, J. B. and D'Addario, K. P. (2001). "The impacts of emoticons on message interpretation in computer-mediated communication." *Social Science Computer Review*. 19.3: 323-345.

Weller, K. (2014). "What Do We Get From Twitter—And What Not? A Close Look At Twitter Research In The Social Sciences." *Knowledge Organization* 41.3: 238-248.

Wells, R. (1956). "Acronymy." In: *For Roman Jakobson. Essays on the Occasion of His Sixtieth Birthday*. Ed. Morris Halle et al. The Hague: Mouton, 662-67.

Williams et al. (2013). "What do people study when they study Twitter? Classifying Twitter related academic papers." *Journal of documentation* 69: 384-410.

Zappavigna, M. (2011). “Ambient Affiliation: A Linguistic Perspective On Twitter.” *New Media & Society* 13.5: 788-806.

INTERNET SOURCES

About Twitter. Twitter, Inc. June 2007. Web. Available at: https://about.twitter.com/company

Best Practices. Twitter, Inc. June 2017. Web. Available at: https://dev.twitter.com/basics

Blagdon, J. (2013). “How emoji conquered the world: The story of the smiley face from the man who invented it.” *The Verge*. 4.3. 2013 Available at: https://www.theverge.com/2013/3/4/3966140/how-emoji-conquered-the-world

Cambridge Dictionary. Web. Available at: http://dictionary.cambridge.org [last accessed 17 July 2011]

Isaac, M. and Ember, S. (2016). “For Election Day Influence, Twitter Ruled Social Media.” *The New York Times*, 8.11. 2016 Available at: https://www.nytimes.com/2016/11/09/technology/for-election-day-chatter-twitter-ruled-social-media.html

McAlone, N. (2016). “These are the most popular apps of 2016 so far.” *Business Insider*, 10.8.2016 Available at: http://www.businessinsider.com/top-apps-of-2016-so-far-2016-8

“Open sourcing Twitter emoji for everyone.” *Twitter Blog*. Twitter, Inc., 6.11. 2014. Web. Available at: https://blog.twitter.com/developer/en_us/a/2014/open-sourcing-twitter-emoji-for-everyone.html

Oxford English Dictionary Online. Web. Available at: www.oed.com [last accessed 17 July 2017]

Twitter, Inc. 2017. Web. Available at: https://twitter.com [last accessed 17 July 2017]

Urban Dictionary. Available at: www.urbandictionary.com [last accessed 17 July 2017]

CONTROL SAMPLE SOURCES

BBC News Daily. (2017). "London fire: Six killed as Grenfell Tower engulfed." *BBC*, 14.6. 2017 Available at: http://www.bbc.com/news/uk-england-london-40269625

Cambers, S. (2017). "Rafael Nadal loses thrilling Wimbledon five-set epic to Gilles Müller." *The Guardian*, 10.11. 2017 Available at: https://www.theguardian.com/sport/2017/jul/10/rafael-nadal-gilles-muller-wimbledon-quarter-finals

Jurejko, J. (2017). "Wimbledon 2017: Rafael Nadal loses to Gilles Muller in 15-13 final set." *BBC*, 10.7. 2017 Available at: http://www.bbc.com/sport/tennis/40562189

Laville, S. et al. (2017). "Grenfell Tower: firefighters search overnight with toll expected to rise." *The Guardian*, 15.6. 2017 Available at: https://www.theguardian.com/uk-news/2017/jun/14/fire-24-storey-grenfell-tower-block-white-city-latimer-road-london

Wilson, J. (2016). "Rafael Nadal knocked out by Gilles Muller in five-set epic." *The Telegraph*, 11.7. 2017 Available at: http://www.telegraph.co.uk/tennis/2017/07/10/rafael-nadal-vs-gilles-muller-wimbledon-2017-live-score-updates/

RESEARCH TOOLS

RStudio Desktop, a freeware software by RStudio, Inc., version 1.0.136. Available at: https://www.rstudio.com

setuptools-36.2.0, a freeware software by Python Software Foundation, version py2.py3. Available at: https://pypi.python.org/pypi/setuptools

Twitter Archiver, a freeware Google add-on by Amit Agarwal, version 20. Available at: https://chrome.google.com/webstore/detail/twitter-archiver/pkanpfekacaojdncfgbjadedbggbbphi [last accessed 16 July 2017]

RESUMÉ

Předkládaná diplomová práce se zabývá procesy zkracování v jazyce sociálních sítí, zejména se zaobírá distribucí zkratek na Twitteru. Jakožto mikroblogovací síť, Twitter dovoluje svým uživatelům vkládat pouze textové příspěvky (tweety) o maximální velikosti 140 znaků, což vede k přirozené tendenci zkracovat jednotlivá slova, avšak i víceslovné výrazy, aby se ušetřilo místo, a tak zvýšil objem zasílané informace. Práce zkoumá hypotézu, že sebraný vzorek 200 zkratek bude rozmanitější a početnější na druhy krácení oproti jiným žánrům, což by mohlo sloužit jako stylistický indikátor tweetového žánru. Dále se předpokládá, že twitterový korpus bude obsahovat typy krácení, jež se nenacházejí v jiných žánrech, které by mohly sloužit jako jeden z určujících, stylisticky příznakových rysů Twitteru. Práce je rozdělená do pěti kapitol.

První kapitola popisuje teoretický podklad pro analýzu. Jelikož se v primárních zdrojích objevují zejména popisy slovotvorného rázu, první část teorie se věnuje třem procesům: mísení (*blending*), mechanickému krácení (*clipping*) a inicialismům (*initialisms*),[15] pod inicialismy jsou zahrnuty i abreviace a akronymy. Mezi primární zdroje patří Bauer a Huddleston (2002), Cannon (1989), Plag (2012) a Quirk a kol. (1985). Dále se v teorii věnuje pozornost internetovým komunitám obecně a sociální síti Twitter konkrétně. Jako poslední část teoretického základu následuje obeznámení s novým lingvistickým oborem, internetovou jazykovědou (*Internet linguistics*), o jehož založení a rozšíření se zasloužil David Crystal (2011). Část popisuje rozdíly mezi tradiční a online komunikací. Na konec jsou zmíněny dostupné studie a jejich poznatky týkající se Twitteru a jevu krácení.

Po teorii následuje metodologická kapitola, která přibližuje parametry sběru vzorku 200 tokenů. Pro jejich extrakci byl zvolen nástroj Twitter Archiver. Aby byl vzorek co nejvíce homogenní, ale zároveň i bohatý na zkratky, byly zvoleny dva trendy, též zvané hashtagy, které určily tematické zaměření (*#GrenfellTower* a *#Wimbledon*). Sebraný vzorek byl dále protříděn podle blíže specifikovaných postupů v metodě a poté z něj bylo vytaženo prvních 100 tokenů zkratek, a to nejdříve z jednoho trendu a až poté z druhého. Takto získaný korpus čítá dohromady 6 540 slov a 433 tweetů. Pro ověření hypotézy bylo rovněž nutné sestavit kontrolní vzorek, aby se s ním twitterový korpus mohl porovnat co se týče četnosti a

[15] Názvy procesů krácení byly přeloženy, aby souhlasily s anglickými protějšky. V české terminologii mohou být názvy procesů odlišné.

rozmanitosti zkratek. Kontrolní vzorek se skládá z 5 článků publikovaných na online portálech britských zpravodajů BBC, The Guardian a The Telegraph. Dva články obsahují reportáž o hořící výškové budově v Londýně, jež byla zachycena v trendu #GrenfellTower, zatímco články o tenisovém utkání mezi Nadalem a Müllerem byly shromážděny tři kvůli jejich krátké délce. Dohromady čítá kontrolní vzorek 6 126 slov.

Ve čtvrté kapitole je prezentována nejdříve kvantitativní analýza, jež nejprve podává výsledky distribuce zkratek v rámci twitterového korpusu a až poté je srovnává s kontrolním vzorkem. V twitterovém korpusu bylo nalezeno 6 různých typů krácení, zatímco v kontrolním vzorku se našly pouze 4. Z analýzy dále vyplývá, že twitterový korpus obsahuje čtyřikrát více zkratek (3,1 %) než kontrolní vzorek (0,8 %), což potvrzuje zkoumanou domněnku, že Twitter je bohatší co do počtu zkratek tak i do počtu metod krácení. Zjistilo se také, že každý třetí tweet obsahuje alespoň jednu zkratku (33,5 %). Oproti tomu bylo zjištěno, že rozmanitost zkratek v obou vzorcích je téměř totožná. Podíl tokenů a lemmat odhalil stejný poměr opakování, 0,44 v twitterovém korpusu a 0,43 v kontrolním vzorku.

Jako nejčastější proces krácení se v obou vzorcích ukázaly inicialismy. Ty v kontrolním vzorku převládaly s 77,6 %, zatímco v twitterovém korpusu netvořily ani poloviční část (43 %). Oba vzorky dále obsahovaly mechanické krácení, komplexní zkratky a logogramy. V žádném se nenašel ani jeden příklad mísení, nevylučuje se však, že by tento proces nebyl na Twitteru produktivní. Je pravděpodobné, že by se mohl najít v trendu s jiným tématem, což by bylo záhodné prozkoumat v nějaké budoucí studii, která by měla vzorek sestaven z více trendů, než jsou zdejší dva.

Při zkoumání inicialismů se zjistilo, že zmiňované dvě subkategorie abreviace a akronymy nejsou vhodné pro propis zkratek v online komunikaci. Jelikož hlavní rozdíl se skýtá v jejich výslovnosti, nedá se posoudit, zda jsou zkratky vyslovené jako celé slovo (akronym) nebo se vyhláskují (abreviace). Jediný pomocný rys pro rozlišení byl, že abreviace krátí jednoslovné i dvouslovné výrazy, zatímco akronymy krátí až tříčlenné a vícečlenné fráze. Při analýze bylo zjištěno, že by bylo vhodné rozlišovat i třetí subkategorii, jež by pokryla jednoslovné výrazy, které se tvoří pomocí spojení prvního písmene nebo prvních dvou písmen určitého slova s posledním písmenem, např. *Mr* z *mister*. Tyto zkratky se vykazují podobnou strukturou a také výslovností, neboť se často vyslovují jako celé původní slovo.

Mimo procesy popsané v primární literatuře, se také našly tři nezmíněné. Pro jejich pojmenování a určení bylo nutné pročíst další zdroje. Bylo zjištěno, že se jedná o

charakteristické znaky jazyka textových zpráv. Jeden z nich, krácení textů formou náhrady slova logogramem, se nacházel i v kontrolním vzorku. Jeho nepřítomnost ve zdrojích je zřejmě důsledek toho, že se nedá mluvit o slovotvorném krácení, nýbrž o substituci. Protože však logogramy plní stejnou funkci jako zkratky, byly do analýzy zahrnuty. Další dva procesy se týkají nestandardního pravopisu (*non-standard spelling*) a vynechaných písmen (*omitted letters*). Oproti inicialismům, mechanickému krácení a logogramům byly spíše krajně zastoupené, jejich přítomnost však byla klasifikována jako stylisticky příznaková pro tweetový žánr. Jako mikroblogovací síť se Twitter totiž inspirovala stručností textových zpráv, s tímto žánrem tedy sdílí stejný stylistický znak.

Celkově se tedy podařilo dokázat, že tweetový žánr je bohatý na počet zkratek a procesů krácení v porovnání s jinými žánry (kontrolním vzorkem). V rámci rozmanitosti zkratek se ukázalo, že twitterový korpus není diversifikovanější, ovšem je možné, že toto zjištění bylo ovlivněno vysokým výskytem zkratky *Rafa*, která se tematicky vázala na *#Wimbledon*. Zkratka *&*, která plnila funkci spojky, byla četná v obou twitterových subkorpusech a proto se nejspíše jedná o obecně rozšířenou zkratku. Podařilo se také ukázat, že tweetový korpus lze rozeznat od jiných žánrů díky dvěma stylisticky příznakovým procesům krácení, a to díky zkratkám psaném nestandardním pravopisem a zkratkám s vynechanými písmeny. Tyto zkratky však nejsou zcela výjimečné pro twitterový žánr, nýbrž jsou společné s žánrem textových zpráv.

APPENDIX

TWITTER SAMPLE

No.	Shortening	Total	GT	W	Meaning	Examples	Method	Specifics
1.	&	33	23	10	and	So proud of the emergency services & community of London. Helpless yet still helping. We will keep praying. #pray #GrenfellTower	logogram	
2.	@	3	1	2	at	What an AMAZING day that was! #wimbledon @ The All England Lawn Tennis Club	logogram	
3.	£	2	2		pound	£10million refurbishment and no fire alarms?? #grenfelltower	logogram	
4.	2	3	3		to	#borisjohnson should hang his head in #shame for his attitude 2 #london #fire #service & #Tory cuts #GrenfellTower	logogram	
5.	4	2	2		for	Substandard fire alarms & flammable materials in this day & age, terrible. #Kensington council r responsible 4 hiring #KCTMO #GrenfellTower	logogram	
6.	am	2	2		ante meridiem = before midday	At 6am NZtime this is what #GrenfellTower looks like. Just a blackened shadow. Cordon still up.	initialism	abbreviation
7.	AO	1		1	Australian Open	It was @andy_murray at #AO and @RafaelNadal today, beaten by left handers using serve and volley. Is this gonna be a thing now? #Wimbledon	initialism	abbreviation
8.	apt	1	1		apartment	All that's left of #GrenfellTower London apt building 18 hrs after fire broke out. A painful reminder visible all around the neighbourhood.	initialism	single word
9.	as	1	1		has	Today as proved again Londoners will always be resilient and help others no matter who you are #GrenfellTower #Londoners	omitted letters	
10.	asap	2	2		as soon as possible	Forest Gate!!!! Donation transport asap #bedsforgrenfell #GrenfellTower	initialism	abbreviation + acronym
11.	av	1	1		have	From SID point of view, theres a need to rethink fire safety of high rise buildings. Suppression av proven fatally inadequate #GrenfellTower	omitted letters	
12.	ave	1	1		avenue	Hi- If you live in Ealing you can drop off donations for #GrenfellTower at *Enchanted* on Northfields ave.	clipping	final
13.	bbc	2	2		British Broadcasting Corporation	If you're a journalist working in London and you're not out there asking tough questions about #GrenfellTower, why are you even there? #bbc	initialism	abbreviation

14.	bcoz	1		1	because	One of those games where u feel the outcome should have been decided by a coin toss bcoz neither deserved to lose #Wimbledon #nadal #muller	non-standard spelling	
15.	bldg	1	1		building	Fire regs & bldg control inspections are not fit 4 purpose. Update needed ASAP. #grenfelltower #BldgRegs #PartB	omitted letters	
16.	BldgRegs	1	1		building regulations	Fire regs & bldg control inspections are not fit 4 purpose. Update needed ASAP. #grenfelltower #BldgRegs #PartB	complex	omitted letters + final clipping
17.	Bro	2	1	1	brother	This must have been a day you and your fearless colleagues have truly been dreading bro #firefighters #GrenfellTower #EmergencyServices	clipping	clipping
18.	BST	1		1	British Summer Time	TUESDAY'S ORDER OF PLAY (Centre Court, from 12.00 BST) Mannarino v Djokovic V.Williams v Ostapenko Konta v Halep #Wimbledon	initialism	abbreviation
19.	CC	5		5	Central Court	Mirka entering that CC stadium like #Federer #Wimbledon	initialism	abbreviation
20.	CET	1		1	Central European Time	So, #Nole plays tomorrow at 13.00 CET! #Wimbledon	initialism	abbreviation
21.	champs	1		1	champions	Watching more former #Aegonilkley champs in action tonight - @MarcusDaniell & Marcelo Demoliner #Wimbledon #comeon! #wishiwasthere	clipping	final
22.	Congrats	4		4	congratulations	#Nadal In such matches there are no losers A class match between 2 gentlemen The way sport should be played Congrats to both #Wimbledon	clipping	final
23.	coz	3	1	2	because	What a match it doesn't matter how good Nadal is coz he just got mullered by muller absolute class match respect to both players #Wimbledon	non-standard spelling	
24.	cray	1		1	crazy	#Wimbledon is cray!	clipping	final
25.	CS	1	1		civil servants	.@foryoubyyou can get emergency payments to CS ... please do share. Occupational funds are here to help! @ACOBenevolence #GrenfellTower	initialism	abbreviation
26.	DM	1	1		direct message	If anyone around Forest Gate has things to donate, DM me I'll arrange pick up. We have a van leaving soon. #GrenfellTower #GlenfellTower	initialism	abbreviation
27.	ETA	1	1		estimated time of arrival	#GrenfellTower - A40 closed both ways (no ETA for re-opening). Heavy traffic on all diversion routes, inc all inputs to Holland Park Rdbt.	initialism	abbreviation
28.	etc	1	1		et cetera = and so on	Lewisham etc - offer of donation transportation #bedsforgrenfell #GrenfellTower	initialism	abbreviation

29.	eu	1	1		European Union	What will be of Britain when @conservatives scrap all #eu regulations? #GrenfellFire #GrenfellTower	initialism	abbreviation
30.	Fab	1		1	faboulous	@Wimbledon What a match! Fab viewing! #wimbledon	clipping	final
31.	gen	1		1	generation	Average age of men's singles q-finalists at #Wimbledon: 1997: 25 years 2008: 26 years 2017: 30 years Where is next gen? @bbctennisnews	clipping	final
32.	gent	1		1	gentleman	What a gent @RafaelNadal is. As gracious in defeat as victory #legend #wimbledon	clipping	final
33.	GMB	1	1		Good Morning Britain	Finally, our so-called PM provides a statement. #disgraceful #Grenfelltower	initialism	abbreviation
34.	hav	1		1	have	Bro, CC is slower! If they were scheduled on CC, Nadal wud hav won in 3/4. #Wimbledon	omitted letters	
35.	hrs	1	1		hours	All that's left of #GrenfellTower London apt building 18 hrs after fire broke out. A painful reminder visible all around the neighbourhood.	omitted letters	
36.	ICYMI	1		1	in case you missed it	ICYMI | The big names all in action on a stunning day of tennis at #wimbledon #7tennis	initialism	abbreviation
37.	inc	1	1		including	#GrenfellTower - A40 closed both ways (no ETA for re-opening). Heavy traffic on all diversion routes, inc all inputs to Holland Park Rdbt.	clipping	final
38.	K	1		1	kilo = thousand	Is Gilles Muller most famous person from Luxembourg after beating Rafa at #Wimbledon? Country has 570K population & not a single one I know.	logogram	
39.	Ken	1	1		Kensington	Used to live in North Ken. Used to work with tenants. Used to be a campaigner highlighting risk of fire. So very sad. #GrenfellTower	clipping	final
40.	KEN&C	1	1		Kensington and Chelsea	#GrenfellTower IS IT BECAUSE PRIME MINISTER U LOST KEN&C TO LABOUR SO WHY SHOULD U BOTHER U HORRIBLE HORRIBLE HUMAN BEING #troysout	complex	clipping + logogram + initialism
41.	KCTMO	3	3		Kensington and Chelsea Tenant Management Organisation	#KCTMO Board Members Anyone spoken out yet? #GrenfellTower #LondonFire	initialism	abbreviation
42.	LBC	1	1		Leading Britain's Conversation	Well said Terry, check this out LBC #GrenfellTower	initialism	abbreviation
43.	libdems	1	1		Liberal Democrats	Was today the best day to announce you were quitting #libdems @timfarron #GrenfellTower	clipping	final compound
44.	LMAO	1		1	Laughing my ass off	Suddenly #Wimbledon looks so soo boring LMAO	initialism	acronym + abbreviation

45.	LOL or lololol	2		2	lots of laugh or laughing out loud	I'm running the #SanFranciscoMarathon in 2 weeks and I know how to inspire myself to the finish line now LOL #RafaNadal #Wimbledon #Nadal	initialism	acronym + abbreviation
46.	mA	1	1		Mashallah = my God	The response from the community has been overwhelming mA! We are taking donations untill tomorrow 3pm! HA1 2SQ #GrenfellTower	initialism	single word
47.	mins	1	1		minutes	The Archbishop of Canterbury @JustinWelby on @BBCLondonNews in a couple of mins, with thoughts on #GrenfellTower fire	clipping	final
48.	mm	2		2	millimetre	#paire looks about 2mm short of a radicalised beard #Wimbledon	initialism	single word
49.	Mr	2	1	1	Mister	Mr @joeottawaystyle at @wimbledon with mrporterlive in his Lock Monaco hat #Wimbledon	initialism	single word
50.	nhs	1	1		National Health Service	Doctor recounts night of Grenfell Tower fire: 'Our first wave of patients came in at 3.45am' #GrenfellTower #nhs	initialism	abbreviation
51.	Ofc	1	1		Of course	Dont make this a race thing'. Ofc it's a race thing. If u need to be told how & why, u're part of the reason it's like this. #GrenfellTower	initialism	abbreviation
52.	OK	1		1	All correct	OK @rogerfederer now that @RafaelNadal is out go and get the trophy. It's yours for the taking now. #Wimbledon	initialism	abbreviation
53.	OMG	4	2	2	oh my god	OMG Gilles Muller stunned Rafael Nadal to wins an epic 5th set 15-13 to and reach his first ever #Wimbledon quarter-final #ScoreBoard	initialism	abbreviation
54.	PM	4	4		Prime Minister	Finally, our so-called PM provides a statement. #disgraceful #Grenfelltower	initialism	abbreviation
55.	pm	2	1	1	post meridiem = past midday	Novak on centre from 12pm tomorrow joke he should had been moved to centre tonight #wimbledon	initialism	abbreviation
56.	PSA	1	1		public service announcement	PSA CB Solutions are opportunistic cocks and I hope they lose business for this. #GrenfellTower	initialism	abbreviation
57.	QFs	2		2	quarter finale	UPSET ALERT // Muller is through to the QFs after a 6-3 6-4 3-6 4-6 15-13 win over Nadal. He will face Cilic next. #Wimbledon	initialism	abbreviation
58.	r	1	1		are	Substandard fire alarms & flammable materials in this day & age, terrible. #Kensington council r responsible 4 hiring #KCTMO #GrenfellTower	logogram	
59.	R	1		1	Round	#RafaelNadal suffers shock defeat in #Wimbledon R4 in nail -biting match of 5 sets against #GillesMuller . #Wimbledon2017 #Upset	initialism	single word

60.	Rafa	21		21	Rafael	Tough luck Rafa and Well done Muller! What a match! #Wimbledon	clipping	final
61.	rbkc	1	1		Royal Borough of Kensington and Chelsea	Fire at Grenfell Tower: Monsoon/Accessorise donate £100,000 via @RBKC #grenfelltower #rbkc #monsoon #fire	initialism	abbreviation
62.	RD	1		1	rough day	Damn RD #Wimbledon	initialism	abbreviation
63.	Rd	1	1		road	Heartbroken by the recent happenings in London. you can drop water/clothes/food @ St Clements Church : 95 Sirdar Rd, W11 4EQ #GrenfellTower	initialism	single word
64.	Rdbt.	1	1		roundabout	#GrenfellTower - A40 closed both ways (no ETA for re-opening). Heavy traffic on all diversion routes, inc all inputs to Holland Park Rdbt.	omitted letters	
65.	Regs	2	2		regulations	Nothing wrong with Building Regs only the implementation of them #GrenfellTower	clipping	fi
66.	rip	1	1		rest in peace in English, from Latin requiescat in pace	My thoughts on the #GrenfellTower catastrophe today... #rip #GlenfellTower #London	initialism	acronym
67.	shld	1	1		should	Who even has a brain that works this way? Who'd vote against smthing tht shld be a basic right?Fucking @Conservatives #GrenfellTower #Tories	omitted letters	
68.	SID	1	1		Security Investigation Division	From SID point of view, theres a need to rethink fire safety of high rise buildings. Suppression av proven fatally inadequate #GrenfellTower	initialism	abbreviation
69.	sm	1	1		so much	it hurts sm more when its so close to home! please please try to donate and give clothes to the help points #GrenfellTower	initialism	abbreviation
70.	smthing	1	1		something	Who even has a brain that works this way? Who'd vote against smthing tht shld be a basic right?Fucking @Conservatives #GrenfellTower #Tories	omitted letters	
71.	SOAS	1	1		School of Oriental and African Studies = University of London	Why SOAS will always be home. #GrenfellTower	initialism	acronym
72.	St	1	1		Saint	Heartbroken by the recent happenings in London. you can drop water/clothes/food @ St Clements Church : 95 Sirdar Rd, W11 4EQ #GrenfellTower	initialism	single word
73.	tho	2	1	1	though	How amazing are our over stretched emergency services tho #GrenfellTower	non-standard spelling	

74.	tht	1	1		that	Who even has a brain that works this way? Who'd vote against smthing tht shld be a basic right?Fucking @Conservatives #GrenfellTower #Tories	omitted letters	
75.	TL	1		1	timeline	Well that's one person on my TL who's happy, Clare Every cloud has a silver lining #Wimbledon	initialism	abbreviation
76.	Tue	1		1	Tuesday	Which is not to say that Rafa wouldn't either but odds on he would stand a better chance after the courts had baked on Tue/Wed #Wimbledon	clipping	final
77.	TV	2		2	television	Thought someone was having an orgasm in this restaurant but it was just #Wimbledon on TV.	initialism	abbreviation
78.	u	9	6	3	you	To those who have fallen in the #GrenfellTower.May you rest in peace to the hero's that keep going,we thank u from the bottom of our hearts	logogram	
79.	UK	2	2		the United Kingdom	UPDATE: #UK: 12 dead as fire engulfs #London tower block - #GrenfellTower	initialism	abbreviation
80.	UPS	1	1		United Parcel Service	Praying for... you know what? Praying for everybody. #Congress #Alexandria #GrenfellTower #London #SanFrancisco #UPS	initialism	abbreviation
81.	US	2		2	the United States	Superb battle at #wimbledon today - shame US #politics couldn't emulate the respect and integrity of the sport.	initialism	abbreviation
82.	v	7		7	versus	#Wimbledon Men's QF after today's play Murray v Querrey Cilic v Muller Raonic v Federer Berdych v Djokovic/Mannarino	initialism	single word
83.	vs	9		9	versus	Feeling on a downer after that Nadal vs Muller match #wimbledon	initialism	single word
84.	w/in	1	1		within	Survivors said the one stairwell to escape w/in #GrenfellTower was allegedly blocked…	complex	omitted letters + logogram
85.	WATTBA	1		1	what a time to be alive	Gilles Müller--Federer's throwback contemporary from the mid-Aughts--just beat Rafa Nadal, who is 4 years younger. #WATTBA #Wimbledon	initialism	abbreviation
86.	Wed	1		1	Wednesday	Which is not to say that Rafa wouldn't either but odds on he would stand a better chance after the courts had baked on Tue/Wed #Wimbledon	clipping	final
87.	WTF or wtf	2	2		what the fuck	Fire went from floor 2 to floor 22 in 15 minutes? WTF? Flammable cladding. Serious questions to answer #grenfelltower	initialism	abbreviation
88.	wud	1		1	would	Bro, CC is slower! If they were scheduled on CC, Nadal wud hav won in 3/4. #Wimbledon	non-standard spelling	
	TOTAL	200	100	100				

CONTROL SAMPLE

No.	Shortening	Total	Meaning	Examples	Method	Specifics
1.	£	3	pound	Grenfell Tower underwent a two-year £10m refurbishment as part of a wider transformation of the estate, that was completed last year.	logogram	
2.	%	2	percent	When I had these last two [match points] I said to myself go for it 100%."	logogram	
3.	9/11	1	the September 11 attacks	The flames, I have never seen anything like it, it just reminded me of 9/11.	logogram	
4.	am	1	ante meridiem = before midday	The first commander on the scene shortly after 1am had been faced with a blaze that spread with a scale and speed greater than he would have anticipated.	initialism	abbreviation
5.	ATP	2	Association of Tennis Professionals	Muller has belatedly begun to realise that potential with two tournaments wins already in 2017 and, seeded 16th, has actually won more matches on grass this summer than anyone on the ATP Tour.	initialism	abbreviation
6.	BBC	6	British Broadcasting Corporation	There must be a "full inquiry" into the fire, newly-elected Kensington MP Emma Dent Coad told the BBC.	initialism	abbreviation
7.	BST	3	British Standard Time	Firefighters, who rescued many people, were called at 00:54 BST and are still trying to put out the fire.	initialism	abbreviation
8.	Dr	1	Doctor	Dr Jim Glocking, technical director of the Fire Protection Association (FPA), an industry body, said a major issue was that insulation underneath cladding on the outside of tower blocks did not need to be fireproof.	initialism	single word
9.	FPA	2	Fire Protection Association		initialism	abbreviation
10.	KCTMO	1	Kensington and Chelsea Tenant Management Organisation	The 24-storey tower, containing about 120 flats, is managed by the Kensington and Chelsea Tenant Management Organisation (KCTMO) on behalf of the council.	initialism	abbreviation
11.	m	3	million	Grenfell Tower underwent a two-year £10m refurbishment as part of a wider transformation of the estate, that was completed last year.	initialism	single word
12.	MGA Autos	1	Marcus and George Antoniades Automobiles	Marco Antoniades, who owns MGA Autos on Latimer Road near Grenfell Tower, said:	complex	abbreviation + final clipping
13.	MP	3	Member of Parliament	Speaking outside the Rugby Portobello Trust emergency centre, the Labour MP said the fire was "absolutely appalling".	initialism	abbreviation
14.	NHS	1	National Health Service	Health Secretary Jeremy Hunt praised the "heroic" response from the emergency services and the NHS hospital staff "working tirelessly to help".	initialism	abbreviation
15.	No.	3	number	The preference was to play the Djokovic v Mannarino match as scheduled on No1 Court.	initialism	single word
16.	pm	4	post meridiem = past midday	Muller even admitted that he feared that the failing light might ultimately prevent a match that had started at 4pm from even being completed.	initialism	abbreviation
17.	Rafa	3	Rafael	I was two sets up, played really well and then Rafa stepped it up.	clipping	final
18.	Rev	1	Reverend	the Rev Mark O'Donoghue, said the church was trying to find hotel rooms	clipping	final

19.	St	5	Saint	St Clement and St James Church was trying to find hotel rooms and bedding for residents of Glenfell Tower.	initialism	single word
20.	UK	2	the United Kingdom	Lessons learnt will be brought out not just across London, but across the UK and globally	initialism	abbreviation
21.	US	1	the United States	"That was tough," said Muller, who has reached his first Grand Slam quarter-final since the 2008 US Open.	initialism	abbreviation
	TOTAL	49				